STATISTICĂ APLICATĂ ÎN ȘTIINȚELE SOCIO-UMANE

Noțiuni de bază - Statistici univariate

Cristian Opariuc-Dan

Constanța, iulie 2008

Fiicei mele, Luana-Ştefana

Statistica, este precum fumatul…

La prima ţigară, ţi-e rău. Dacă continui, începe să-ţi placă şi în final termini prin a fi dependent.

Diferenţa dintre statistică şi fumat este că, statistica nu dăunează grav sănătăţii.

Cuprins

Cuprins 3
Cuvântul autorului 7
I. Natura măsurării în ştiinţele socio-umane 11
I.1 Variabile 14
I.1.1 Variabile discrete 16
I.1.2 Variabile continui 17
I.2 Scale (nivele) de măsurare 18
I.2.1 Scale neparametrice 19
I.2.2 Scale parametrice 22
I.3 Prezentare generală SPSS for Windows 26
I.3.1 Bazele de date – creare, salvare, deschidere 31
II. Organizarea datelor 54
II.1 Sistematizarea datelor 56
II.2 Reprezentarea grafică a datelor 63
III. Statistici descriptive – Tendinţa centrală 67
III.1 Populaţie şi eşantion 68
III.1.1 Eşantionarea simplu randomizată 70
III.1.2 Randomizarea pe cote 71
III.1.3 Eşantionarea stratificată 72
III.1.4 Eşantionarea pe cluster 73
III.2 Indicatori ai tendinţei centrale 73
III.2.1 Media 75
III.2.2 Mediana 80
III.2.3 Modul 83

III.2.4 Precizia indicatorilor tendinței centrale 85
III.2.5 Obținerea indicatorilor tendinței centrale în SPSS 88
IV. Media populației. Reprezentări grafice ale tendinței centrale 96
IV.1 Reprezentări grafice 97
IV.2 Analiza grafică a tendinței centrale 107
IV.2.1 Graficul „tulpină şi frunze” 108
IV.2.2 Graficul „cutie” (box-plot) 111
IV.3 Tratarea scorurilor extreme 115
IV.3.1 Obținerea graficelor pentru tendința centrală în SPSS 117
V. Statistici descriptive – Împrăştierea 122
V.1 Indicatori elementari ai împrăştierii 124
V.1.1 Amplitudinea de variație 124
V.1.2 Abaterea cuartilă sau abaterea intercuartilă 125
V.2 Indicatori sintetici ai împrăştierii 129
V.2.1 Diferența medie Gini 129
V.2.2 Oscilația 130
V.2.3 Abaterea medie 133
V.2.4 Abaterea mediană 135
V.2.5 Dispersia 135
V.2.6 Abaterea standard 138
V.2.7 Coeficientul de variație 141
V.3 Obținerea indicatorilor împrăştierii în SPSS 142
VI. Distribuția normală 146
VI.1 Modalitatea 148
VI.2 Simetria 149
VI.2.1 Coeficientul Yule 150
VI.2.2 Coeficientul de asimetrie Fisher 150

VI.3 Boltirea sau excesul 155
VI.4 Caracterizarea distribuției normale 159
VI.4.1 Notele „z" şi note standard derivate 161
VI.5 Transformarea datelor brute 168
VI.6 Sisteme de etalonare 172
VI.6.1 Etaloane în cuantile 177
VI.6.2 Etaloane normalizate 180
VI.7 Analiza distribuției şi realizarea etaloanelor în SPSS 186
VI.7.1 Calculul notelor „z" în SPSS 200
VII. Probabilități şi semnificație statistică 206
VII.1 Patru reguli de probabilitate 209
VII.1.1 Probabilitatea simplă pentru evenimente egale 209
VII.1.2 Evenimentele mutual exclusive 211
VII.1.3 Evenimentele independente 213
VII.1.4 Probabilitatea condiționată 215
VII.2 Eroare standard şi intervale de încredere 216
VII.3 Metoda ştiințifică în ştiințele socio-umane 224
VII.3.1 Enunțarea problemei 225
VII.3.2 Formularea ipotezelor 227
VII.3.3 Proiectarea cercetării 243
VII.3.4 Efectuarea observațiilor 244
VII.3.5 Interpretarea datelor 244
VII.3.6 Formularea concluziilor 246
VIII. Eşantionare şi reprezentativitate 249
VIII.1 Populația 250
VIII.2 Surse de date 251
VIII.2.1 Sursele principale de date 251

VIII.2.2 Sursele secundare de date 252
VIII.3 Eşantionul 252
VIII.3.1 Mărimea eşantionului 255
VIII.3.2 Caracteristicile populației 256
VIII.3.3 Procedee de eşantionare 256
VIII.3.4 Stabilirea eşantionului prin analiză 258
VIII.3.5 Determinarea mărimii eşantionului 276
VIII.4 Construcția eşantioanelor în SPSS 283
VIII.4.1 Crearea unui eşantion în SPSS 284
VIII.4.2 Analiza eşantionului 300
VIII.4.3 Investigarea proprietăților eşantionului 306
Bibliografie 317
În loc de încheiere 319

Cuvântul autorului

Bun venit la cursul de statistică aplicată în ştiinţele socio-umane. Dacă aţi cumpărat această carte, probabil că sunteţi student sau absolvent al unei facultăţi pe care aţi urmat-o tocmai din dorinţa de a scăpa de matematică, de calcule, probabil că vă plac mai mult cuvintele, vă place să vorbiţi cu oamenii, să-i înţelegeţi şi să-i ajutaţi, să pătrundeţi fenomenele şi evenimentele sociale. Îmi imaginez surprinderea şi îngrijorarea dumneavoastră în momentul în care, abia păşind pe băncile facultăţii, aţi aflat cu stupoare că trebuia să studiaţi… statistică. Mai mult, că statistica v-a urmat în toţi anii de facultate, inclusiv la licenţă. Ştiu! Şi eu la rândul meu am trăit aceleaşi sentimente şi nu mă pot lăuda că am obţinut, de la început, performanţe deosebite în acest domeniu. Asta pentru că statistica nu se lasă uşor dezvăluită. Totuşi, statistica reprezintă fundamentul ştiinţific al unor discipline precum psihologia, sociologia, antropologia şi multe altele. Fără sprijinul său, nu am reuşi să vorbim astăzi de inteligenţă, trăsături de personalitate, tipuri de memorie, nu am avea teste care să le măsoare la nivelul la care reuşesc acum să le măsoare.

Statistica nu este atât de grea. Nu este nevoie să ştim concepte avansate de matematică superioară pentru a înţelege şi aplica principii statistice. Dacă ştiţi să adunaţi, să scădeţi, să înmulţiţi şi să împărţiţi, sunt sigur că veţi putea învăţa foarte bine şi statistica. Nu trebuie să vă sperie formulele complexe şi neinteligibile. De cele mai multe ori, aceste formule nu sunt altceva decât algoritmi, paşi simpli, proceduri de calcul elementar, care vă introduc în lumea fascinantă a probabilităţilor. Chiar dacă uneori apar termeni precum logaritmi, integrale, derivate şi limite, staţi liniştiţi. Nimeni nu vă pune să le calculaţi. Unele calcule pot presupune utilizarea unui calculator ştiinţific şi… cam atât.

Am încercat să fac din acest volum un material special şi am evitat de fiecare dată să utilizez concepte matematice laborioase. Fiecare formulă pe care o prezint aici va fi explicată, descompusă. Uneori abundenţa de exemple

şi repetiţiile sunt destinate fixării unor concepte mai dificile sau mai importante. Sigur că statistica presupune încă de la început stăpânirea unor concepte noi. Nu putem avansa dacă nu înţelegem sensul termenului de a măsura sau dacă nu ştim ceea ce este o variabilă, un nivel de măsură sau un indicator statistic. Mulţi dintre dumneavoastră abia acum iau contact cu aceşti termeni şi, desigur, lucrul acesta îi sperie. Fiţi fără grijă. O mare parte din terminologia statistică o folosiţi deja şi aţi folosit-o de mii de ori. Doar că nu ştiaţi că se numeşte aşa şi nici de unde provine. Împreună vom face lumină.

Acest volum este unul practic. Fiecare capitol îşi propune obiective precise şi abundă de exemple şi exerciţii concrete. În definitiv, mă interesează să stăpâniţi practica statisticii, să o utilizaţi în activitatea dumneavoastră de zi cu zi şi nu să facem speculaţii teoretice pe această temă. Majoritatea capitolelor sunt formate din două părţi: un fundament teoretic şi conceptual însoţit de algoritmii (paşii) clasici ai unui procedeu statistic şi o aplicaţie folosind un pachet de programe de prelucrări statistice, SPSS for Windows. În această carte am utilizat mai multe versiuni SPSS, ultima fiind versiunea 15. Sunt sigur că principiile prezentate aici vor putea fi aplicate şi în versiunile următoare. La momentul scrierii acestei cărţi, ultima versiune de SPSS este versiunea 16. În general, fiecare capitol se finalizează cu un set de întrebări şi exerciţii pe care vă invit să le parcurgeţi singuri, deoarece sunt menite să stabilizeze cunoştinţele dobândite.

Totuşi, de ce este important să studiem statistica? Desigur, argumentele sunt nenumărate, însă mă voi rezuma doar la câteva. Cel ce nu cunoaşte statistică nu poate înţelege articolele de specialitate. Toate cercetările care se fac în acest domeniu folosesc metode statistice mai mult sau mai puţin elaborate. Un specialist în domeniu care nu cunoaşte statistică, nu poate citi şi înţelege în profunzime conţinutul unui material de specialitate, corectitudinea sau incorectitudinea datelor expuse. În altă ordine de idei, necunoscând statistică, nu se poate face cercetare experimentală, nu pot fi înţelese principiile constructive ale unui instrument sau metodă de măsurare a unor calităţi psiholo-

gice sau a unor fapte sociale şi, prin urmare, un asemenea instrument nu va putea fi utilizat corect.

Materialele pe care le-am folosit pentru a scrie această carte sunt, în marea lor majoritate, prezentate în bibliografia finală. De asemenea, am utilizat multiple surse de pe Internet, precum şi experienţa mea ca practician şi dascăl în ale statisticii. Nu pot să nu-i menţionez pe cei care mi-au fost primii îndrumători în această aventură şi de la care am avut şi încă mai am de învăţat foarte mult. Astfel, îi sunt recunoscător profesorului doctor Cornel Havârneanu, mentorul şi prietenul meu care mi-a ghidat formarea şi m-a susţinut permanent, prietenului meu asistent doctor Ovidiu Lungu, un statistician înnăscut, care m-a determinat să înţeleg că acesta ar fi domeniul ce mi s-ar potrivi cel mai bine şi care mi-a sprijinit începuturile, lector doctor Loredana Ruxandra Gherasim, cu care de nenumărate ori am învăţat să aplic practic ceea ce ştiam teoretic sau să îmi fundamentez teoria plecând de la practică, profesor universitar doctor Aurel Stan, care a avut bunăvoinţa şi răbdarea, alături de cei menţionaţi mai sus, să îmi citească lucrarea şi să îmi facă observaţii pertinente pe seama ei şi prin intermediul căruia am luat cunoştinţă de statistică prima dată, în anii studenţiei, care ulterior mi-a furnizat informaţii valoroase, ce mi-au permis să mă perfecţionez. Fără contribuţia lor, în mod sigur nu aş fi ajuns niciodată să scriu aceste rânduri şi lor le datorez recunoştinţa mea.

Nu pot, de asemenea, să nu mulţumesc unei… situaţii, situaţia care m-a determinat să vin din Iaşi la Constanţa. Am învăţat că… statistic un eşec personal creează premisele unei relansări. Dacă această situaţie, dacă propriile mele eşecuri nu ar fi existat, probabil că această carte nu ar fi existat nici ea. Întotdeauna, chiar atunci când ceva pare sigur imposibil, să ştiţi că soluţia e chiar la îndemâna dumneavoastră. Referitor la acest fapt, nu pot să nu mulţumesc fiicei mele, Luana-Ştefana, care mi-a fost sprijin şi speranţă în toate momentele dificile şi care în ciuda vârstei nu a încetat să mă sprijine şi să mă încurajeze. Îi mulţumesc şi îi dedic prima mea lucrare.

Această carte nu ar fi văzut tiparul fără susţinerea profesorului universitar doctor Mircea Miclea, care m-a încurajat, m-a înţeles şi a sprijinit apariţia acestui volum. Îi mulţumesc şi îmi exprim recunoştinţa şi pe această cale. Nu pot, de asemenea, să nu îi amintesc pe părinţii mei şi pe toţi prietenii mei, mai mult sau mai puţin apropiaţi, datorită cărora am avut energia să încep această lucrare şi cărora le mulţumesc pentru că m-au ajutat să trec peste momentele dificile din ultimii ani şi faţă de care nu pot să nu îmi exprim recunoştinţa.

Cât aş încerca, nu reuşesc să îmi scot din minte vorbele şi imaginea primului meu profesor adevărat de matematică, domnul Crezant Ghenghea, din cadrul liceului industrial (pe atunci), actualmente Colegiu Naţional „Nicu Gane" din Fălticeni, care a reuşit să-mi inducă, cum nu a făcut nimeni, dragostea faţă de această disciplină. Chiar dacă nu îl înţelegeam în acel moment al adolescenţei mele, el a fost primul (şi singurul) care m-a văzut cu aptitudini în acest domeniu, aptitudini pe care eu însumi mi le repudiam. Au trebuit să treacă mulţi ani ca să ajung la vorbele sale: *„Eşti pentru ştiinţe exacte. Nu poţi scăpa de ceea ce eşti. Ai să-mi dai dreptate mai târziu"*. Şi i-am dat dreptate. Mi-aş dori să fie acum mândru de ce-am realizat şi voi folosi această ocazie pentru a-i mulţumi.

Nu în ultimul rând, vă mulţumesc dumneavoastră, care aţi cumpărat această carte şi vă invit să mă urmaţi prin călătoria în lumea statisticii, să vă convingeţi că lucrurile sunt mai simple decât credeţi. Eu nu pot decât să vă garantez că vă voi scoate la liman. Observaţiile dumneavoastră sunt foarte importante pentru mine şi mă pot orienta în viitoarele lucrări sau reeditări ale acestui volum. Dacă unele informaţii nu vi se par clare, dacă doriţi şi altceva, nu vă rămâne decât să-mi scrieţi pe adresa copariuc@gmail.com şi să-mi comunicaţi opinia dumneavoastră. Vă mulţumesc şi vă doresc succes!

Constanţa, duminică, 13 iulie 2008

I. NATURA MĂSURĂRII ÎN ŞTIINȚELE SOCIO-UMANE

În acest capitol se va discuta despre:

- Sensul şi semnificaţia termenului de măsurare;
- Variabile discrete şi continui;
- Niveluri de măsurare;
- Descrierea programului SPSS;
- Crearea unei baze de date SPSS;
- Operaţii elementare în SPSS.

După parcurgerea capitolului, cititorii vor fi capabili să:

- Înţeleagă conceptul de măsurare în ştiinţele socio-umane;
- Distingă între variabilele continui şi variabilele discrete;
- Diferenţieze nivelurile de măsurare;
- Se familiarizeze cu aplicaţia SPSS;
- Creeze o bază de date SPSS şi să definească corect variabilele.

Statistica se referă la măsurare şi probabilităţi. Cuvântul statistică, provine din latinescul „*statista"* şi desemna, pe vremuri, persoanele care se ocupau cu afacerile statului, care numărau populaţia sau realizau alte acţiuni ce ajutau statul să gestioneze mai bine politica de taxe şi costurile războaielor. În perioada medievală, prin statistică, armatorii îşi calculau costurile echipării corăbiilor, incluzând în calculele lor şi probabilitatea ca acestea să fie atacate de piraţi sau de a naufragia (Lungu, 2001).

Foarte multe dintre conceptele statistice, au avut rădăcini în alte discipline, unele complet diferite de matematică. Astfel, corelaţiile provin din… biologie, din analiza asemănări dintre copii şi părinţii lor. Analiza de varianţă provine din fabricile de bere şi se folosea la alegerea tipului potrivit de orz şi a timpului optim de fermentare, astfel încât berea să aibă un anumit gust. Mai interesant este faptul că însăşi teoria măsurării îşi are originea în psihologie, mai precis în studiul inteligenţei şi al personalităţii umane iar testele nepara-

metrice provin din sociologie. Unii statisticieni încercau chiar să demonstreze existența lui Dumnezeu cu ajutorul numerelor iar astăzi, statistica joacă un rol important în viața noastră, fie că ne referim la psihologie, sociologie, ştiințe economice sau alte discipline.

În viziunea lui T. Rotariu, „statistica se axează în principal pe tratarea informațiilor numerice obținute la nivelul unor mulțimi de entități, informații prelevate de la fiecare entitate în parte (sau de la o submulțime) şi care conduc la rezultate cu referință la ansamblu, şi nu la entitățile componente luate individual." (Rotariu, şi alţii, 2006)

Operațiile statistice presupun, prin urmare, un proces de măsurare. La o analiză atentă, sintagma „a măsura un obiect" nu înseamnă altceva decât a măsura *proprietățile* unui obiect. Noi nu putem măsura obiectele, ci, pentru a le cunoaşte, măsurăm anumiți indicatori ai proprietăților acestora. Este adevărat că, în anumite ştiințe (fizica, chimia etc.), măsurarea dobândeşte un caracter mai concret, deoarece obiectele şi proprietățile acestora sunt mai accesibile observării directe. Putem măsura cu uşurință lungimea şi lățimea unei mese (deci indicatori ai proprietăților lungime şi lățime a obiectului masă), deoarece aceste două proprietăți sunt direct accesibile observației noastre. Însă, în momentul în care discutăm de anxietate, depresie, inteligență, atenție, responsabilitate, nivel de trai etc. constatăm că aceste proprietăți nu pot fi direct observate. Prin urmare, va trebui să deducem aceste caracteristici din observarea unor indicatori presupuşi ai proprietăților. **Aşadar, măsurarea în ştiințele socio-umane are un caracter subtil, care scapă observării directe.** (Vasilescu, 1992)

În ştiințele socio-umane noi nu cunoaştem şi uneori nu putem cunoaşte în mod precis semnificația mărimilor pe care le determinăm, iar cifrele pe care le obținem sunt aproximări mai mult sau mai puțin exacte ale realității. Acest lucru se întâmplă datorită apariției erorilor de măsurare şi datorită faptului că realitatea socială nu are un caracter fix, ci unul fluctuant, variabil.

Cantitatea în care un obiect este saturat într-o caracteristică, nu este reprezentată printr-un punct, ci printr-un nor de puncte, care oscilează într-un anumit interval, în jurul unei valori. (Vasilescu, 1992)

Discutând despre sensul termenului de „a măsura", acceptăm definiția dată de *S. S. Stevens (1959)* prin care, **„în sensul său larg, măsurarea înseamnă atribuirea de numere obiectelor sau evenimentelor, potrivit unor reguli."** Măsurarea este deci, o funcție, prin care unui obiect îi corespunde un număr şi numai unul, la o **anumită** măsurare. Această funcție de atribuire va trebui să fie **suficient de clară,** pentru a permite ca unui obiect să i se atribuie un număr şi numai unul şi să permită să decidem concret şi corect, ce număr va fi atribuit fiecărui obiect, dar şi **suficient de simplă** pentru a putea fi aplicată. (Vasilescu, 1992)

Spre exemplu, codificarea genului biologic al subiecților, este o măsurare; dacă persoana este bărbat, i se acordă valoarea 1 iar dacă persoana este femeie, i se acordă valoarea 2. Observăm că regula de atribuire nu permite confuzii – bărbatul este notat cu 1 iar femeia cu 2 – oferă o selecție exhaustivă de categorizare a oricărui element care apare în mulțimea „gen biologic" şi este suficient de simplă pentru a fi utilizată. Acesta este un proces de măsurare conform definiției lui Stevens, deşi putem constata cu uşurință lipsa posibilității de ierarhizare. Nu putem construi, la acest nivel, o ierarhie a subiecților, deoarece nivelul de măsurare este unul categorial. Despre nivelurile şi scalele de măsurare, vom discuta însă mai târziu.

Un alt exemplu de măsurare îl reprezintă codificarea gradelor didactice: notăm cu 0 preparatorul, cu 1 asistentul, 2 – lectorul, 3 – conferențiarul, 4 – profesorul. Observăm, din nou, univocitatea şi simplitatea regulii; într-o universitate, fiecare cadru didactic poate să facă parte dintr-o categorie şi numai din una iar fiecare persoană se află, la un moment dat, într-o asemenea categorie. În plus, observăm existența unei ierarhii; conferențiarul este inferior în grad profesorului şi superior în grad lectorului, fără însă a se putea pre-

ciza nici cu cât nici de câte ori. Aceasta este însă, din nou, o problemă a nivelului de măsurare, pe care o vom aborda ulterior.

I.1 Variabile

Am stabilit deja faptul că obiectele pot fi cunoscute prin măsurarea indicatorilor proprietăţilor sale. Aşadar, obiectele devin măsurabile prin stabilirea caracteristicilor lor. **Caracteristica este, prin urmare, o particularitate, o însuşire a unui obiect sau fenomen, care constituie obiectul măsurării**. De exemplu, o masă poate fi caracterizată prin lungime, lăţime, înălţime, greutate, formă, culoare etc. Toate acestea constituie caracteristici prin care încercăm să descriem cât mai exact obiectul măsurat. Cu cât avem mai multe asemenea caracteristici, cu atât obiectul se conturează mai precis.

Caracteristicile prin care obiectul este descris, poartă numele de **variabile.** O variabilă reprezintă un concept cheie în statistică şi nu este altceva decât **un nume pentru un element a cărui principală proprietate este aceea că variază, îşi modifică valorile.** Prin această proprietate principală, variabilele se disting de **constante,** elemente care au valori fixe. De exemplu, într-un studiu efectuat pe o populaţie generală, genul biologic poate fi o variabilă dacă lotul de cercetare cuprinde atât bărbaţi cât şi femei. Dacă însă intenţionăm să desfăşurăm o cercetare numai pe femei, atunci genul biologic devine o constantă, deoarece nu prezintă proprietatea principală a variabilelor, aceea de a-şi modifica valorile.

Modalitatea de realizare a unei variabile (modul în care îi sunt atribuite valorile) constituie un eveniment întâmplător (probabilistic), de aceea variabila se mai numeşte şi **variabilă aleatoare sau variabilă stocastică.** De exemplu, înălţimea unor elevi dintr-o clasă este o variabilă care poate lua aleatoriu diferite valori pe care le pot avea elevii la această caracteristică. Desigur, valorile pe care le poate lua variabila se grupează într-un anumit interval de valori. Nu putem vorbi de elevi de 5 centimetri, după cum nu putem găsi elevi de 5 metri. De aceea, putem considera că înălţimea elevilor din

clasa a X-a este o variabilă stocastică ce poate lua aleatoriu valori din **domeniul de definiție** 150 – 190 centimetri.

Aşadar, elementele de bază care compun domeniul de definiție al unei variabile aleatorii poartă numele de **valori** sau **scoruri**. În momentul în care aceste elemente sunt numere, variabila se mai numeşte şi **variabilă alea-numerică.**

O variabilă aleatorie exprimă, aşadar, variația unei caracteristici. După cum am spus, înălțimea elevilor este o variabilă alea-numerică; la fel greutatea acestora, capacitatea toracică, temperatura corpului sau numărul de la pantofi. Nu toate variabilele sunt însă alea-numerice. Culoarea ochilor, culoarea părului, genul biologic, tipul temperamental etc. sunt variabile ale căror valori nu pot fi exprimate prin numere, deşi sunt şi ele variabile aleatorii. Asemenea variabile nu sunt considerate variabile alea-numerice.

Toate aceste caracteristici pot fi măsurate, înregistrate şi catalogate şi de asemenea diferă de la o persoană la alta sau de la o situație la alta. De aceea, toate aceste concepte le vom include sub denumirea generică de **variabile**.

De ce suntem însă atât de interesați de aceste variabile, de ce sunt importante şi la ce ne folosesc? Răspunsul este acela că, nu putem fi mulțumiți doar cu ideea că variabilele… variază. Ne interesează, în principiu, să înțelegem de ce variază, când şi în ce condiții variază, care este efectul acestor variații ale lor. Pentru aceasta, variabilele vor trebui mai întâi definite, apoi înregistrate şi în final supuse analizei, creând premisele extragerii unor concluzii şi, implicit, a generalizării.

Realizarea unei variabile prin intermediul scorurilor creează posibilitatea abordării conceptului statistic de **frecvență**. Frecvența reprezintă răspunsul la întrebarea „câți indivizi, câte cazuri populează fiecare categorie a

variabilei" sau, aşa cum exprimă T. Rotariu, „numărul de indivizi statistici care populează fiecare clasă a caracteristicii" (Rotariu, şi alţii, 2006).

Dacă ne propunem să înregistrăm persoanele dintr-o sală de curs sub aspectul culorii ochilor vom avea, spre exemplu, 20 de persoane cu ochi albaştri, 13 persoane cu ochi verzi, 15 persoane cu ochi negri şi 50 de persoane cu ochi căprui. Valorile care însoţesc categoriile variabilei „culoarea ochilor" reprezintă tocmai frecvenţele – mai exact **frecvenţele absolute** deoarece mai pot exista frecvenţe exprimate procentual (frecvenţe relative) şi frecvenţe cumulate, despre care vom discuta în următorul capitol.

O variabilă aleatoare poate fi **discretă** sau **continuă,** după cum mulţimea de definiţie a valorilor sale poate fi numărabilă (discontinuă) sau nenumărabilă (continuă)[1].

I.1.1 Variabile discrete

În cazul variabilelor discrete (care, în funcţie de nivelul de măsurare pot fi categoriale sau ordinale – vom vedea imediat ce înseamnă acest lucru) **mulţimea de definiţie a valorilor sale este o mulţime discontinuă** (elementele sale pot fi numărate) iar între două valori nu poate fi interpusă, *în permanenţă*, o valoare intermediară.

Spre exemplu, variabila „culoarea ochilor" este o variabilă discretă; ea poate lua valorile: negru, căprui, verde, albastru şi putem clasifica subiecţii într-una din aceste patru categorii. Iată că, mulţimea de definiţie a valorilor aceste variabile este una numărabilă (are doar patru valori) şi nu avem posibilitatea ca între două valori să includem, *în permanenţă***,** o a treia valoare intermediară. Nu putem vorbi despre ochi verde-albastru şi apoi verde-verde-albastru şi aşa mai departe, la infinit.

[1] Există mai multe clasificări ale variabilelor. Le vom aborda atunci când situaţia o va impune pentru a nu complica inutil, la acest nivel, expunerea noastră.

Un asemenea exemplu de variabilă este o **variabilă discretă nominală** sau **variabilă discretă categorială,** deoarece permite doar clasificări, fără a putea vorbi de o relație de ordine sau ierarhie între valorile pe care le poate lua respectiva variabilă.

Un alt exemplu de variabilă discretă este variabila „grad didactic". Ea poate lua valorile: preparator, asistent, lector, conferențiar, profesor. Sigur că, şi în acest caz putem clasifica subiecții în funcție de gradul lor didactic. Nu putem însă vorbi, nici acum, de o nouă categorie intermediară, de exemplu asistent-lector şi apoi asistent-lector-lector şi aşa mai departe, lucru care îi conferă caracterul discret. De data aceasta însă, putem pune o relație de ordine între valorile variabilei. Putem vorbi de faptul că asistentul este inferior în grad lectorului sau profesorul este superior în grad conferențiarului. O asemenea variabilă se numeşte **variabilă discretă ordinală,** sau **variabilă discretă de rang.**

I.1.2 Variabile continui

În cazul variabilei continui, mulțimea de definiție a valorilor sale este una continuă, ne-numărabilă. Aceste variabile se mai numesc şi **variabile scalare** (de interval sau de raport). În această situație, între două valori ale variabilei continui putem găsi, oricând, o valoare intermediară. Cu alte cuvinte, o caracteristică continuă poate lua orice valoare între anumite limite.

Spre exemplu, variabila „înălțimea subiecților" este o variabilă de tip scalar (continuu). Între un subiect de 178 centimetri şi un subiect de 180 centimetri, putem găsi, *oricând*, un subiect de 179 centimetri, apoi între unul de 178 şi unul de 179 putem găsi un subiect de 178,5 centimetri şi aşa mai departe, la infinit.

O asemenea variabilă se caracterizează prin faptul că, tipul de date măsurat de aceasta este format din unități de dimensiune egală. Ca exemplu de variabile continui putem da greutatea subiectului, intervalul de tip petrecut

pentru a juca jocuri agresive pe computer etc. Astfel, înălţimea subiecţilor o putem măsura în centimetri, greutatea în kilograme, timpul în minute sau ore, toate acestea făcând referire la un înalt nivel de precizie, cu valori zero absolute şi posibilităţi de construire a scalelor de raport – putem spune, de exemplu, că un subiect este de două ori mai greu decât altul sau de două ori mai înalt.

I.2 Scale (nivele) de măsurare

În capitolul anterior am văzut că variabilele discrete pot fi categoriale sau de rang; cele continui – de interval sau de raport. Dacă acceptăm definiţia lui Stevens, atunci putem intui că măsurarea poate fi realizată la diferite niveluri de măsurare, organizate ca în figura de mai jos.

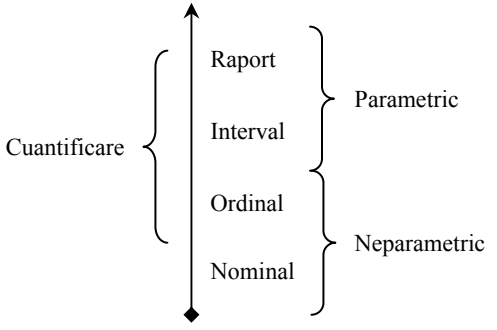


Figura 1.1 – Reprezentarea schematică a nivelurilor de măsurare

Nivelul nominal de măsurare (clasificare) şi nivelul ordinal (de rang sau ierarhic) se grupează în **scala neparametrică (nonmetrică)** de măsurare.

Nivelul de interval (intervale egale) şi nivelul de raport (proporţii) formează **scala parametrică (metrică)** de măsurare.

Observăm că **termenul de cuantificare** dobândeşte sens începând cu scala ordinală, deoarece abia la acest nivel ne permitem realizarea unor ierarhii. Scala nominală nu acceptă cuantificare ci doar clasificare.

Se cuvine să facem menţiunea că, proprietăţile scalelor sunt incrementale, adică proprietăţile unui nivel inferior de măsurare sunt valide şi la un nivel superior de măsurare, însă nu şi reciproc.

I.2.1 Scale neparametrice

Grupează scala nominală şi scala ordinală, scale la nivelul cărora putem pune, cel mult, o relație de ordine între nivelurile variabilei. Nu putem vorbi de parametri şi nici de statistici metrice. Nu are sens calculul mediei sau al altor indicatori din sfera metrică, deoarece nu putem stabili un interval (cu cât) față de care o valoare este mai mare în comparație cu alta.

I.2.1.1 Scala nominală (de clasificare)

Reprezintă primul nivel de măsurare şi constă în clasificarea obiectelor în funcție de existența sau inexistența unei caracteristici. Această clasificare presupune existența unor categorii disjuncte, astfel încât fiecare obiect să-şi găsească locul într-o categorie şi numai în una (Vasilescu, 1992).

Pentru ca scala să aibă sens, trebuie să existe cel puțin două categorii. O variabilă nominală cu două categorii poartă numele de **dihotomie** sau **clasificare dihotomică**. Dacă există mai multe categorii, vorbim despre **polihotomii** sau **clasificări multiple (polihotomice).** Împărțirea subiecților în bărbați şi femei reprezintă o clasificare dihotomică. Împărțirea în funcție de profesie, culoarea ochilor sau a părului, grupa sanguină etc. sunt tot atâtea exemple de polihotomii.

În mod curent folosim această scală în vederea categorizării subiecților în funcție de unul sau mai multe criterii (de obicei în cercetarea psihologică sau sociologică) sau atunci când culegem date în urma anchetei sau observației.

Proprietăți ale scalei:

- Transformările permise la nivelul acestei scale sunt (Vasilescu, 1992):
    - Redenumirea – atribuirea unor noi nume (în loc de masculin/feminin putem spune bărbat/femeie);

  - Permutarea – schimbarea ordinii elementelor, deoarece nu se poate stabili o ierarhe între valorile unei variabile pe această scală. Această proprietate este caracteristică doar scalei nominale.

- Operaţii statistice permise – în principal operaţiile de bază, derivate din numărare, astfel:
  - Frecvenţa absolută şi relativă (procentul), câte observaţii au fost incuse în fiecare categorie (de exemplu 25 de bărbaţi şi 40 de femei sau 30% bărbaţi şi 78% femei);
  - Valoarea modală (modul), categoria cu frecvenţa cea mai mare (de exemplu din 200 de subiecţi, 150 au ochi albaştri, aceasta fiind categoria cu frecvenţa cea mai mare – modul);
  - Verificarea statistică prin χ^2 – procedeu de comparare a frecvenţelor şi de verificare dacă diferenţele dintre ele sunt întâmplătoare sau, din contra, semnificative, neîntâmplătoare;
  - Coeficienţi de corelaţie între două variabile cu două valori fiecare, coeficientul φ, coeficientul tetragoric, coeficientul de contingenţă, toate bazându-se pe lucrul cu frecvenţe absolute sau relative.

I.2.1.2 Scala ordinală (de rang sau topologică)

Reprezintă al doilea nivel de măsurare şi constă în ierarhizare, în funcţie de mărimea unei caracteristici, fără însă a se putea preciza cu cât un nivel ierarhic este superior sau inferior altuia şi nici de câte ori.

Spre deosebire de scala nominală (ale cărei proprietăţi le include), scala ordinală permite stabilirea unei relaţii de ordine între date.

Strict vorbind, acesta este nivelul la care ne situăm în ştiinţele socio-umane în general şi în psihologie în particular. Din perspectiva statisticii, ca disciplină matematică, nu putem vorbi de medie, abatere standard sau alți indicatori care presupun un nivel de măsură cel puțin de interval, deoarece o asemenea scală nu există în domeniul socio-uman la o rigoare (încă) demonstrabilă matematic. Scorurile testelor de inteligență, aptitudini, personalitate sunt variabile având valori simplu ordonate la un nivel ordinal de măsură. Un test de inteligență nu ne arată, de fapt, cantitatea de inteligență a subiecților, ci ierarhizarea acestora, poziția, rangul pe care îl ocupă fiecare într-o populație.

Lansarea, în 1905, a scalei metrice a inteligenței de către Alfred Binet, se făcea cu următoarea mențiune din partea autorului: *„calitățile intelectuale nu se măsoară precum lungimile, nu satisfac cerința aditivității; scala permite un clasament ierarhic între inteligențe diferite şi, din raţiuni practice, acest clasament echivalează cu o măsurare."* (Radu, şi alţii, 1993). Autorul a sesizat foarte bine limitele măsurării în ştiinţele socio-umane încă din acea perioadă, limite pe care mulți psihologi actuali astăzi le ignoră. Nu ne miră faptul că, matematicienii consideră aplicarea statisticii în psihologie ca fiind defectuoasă.

Datorită faptului că, dacă ne-am limita doar la o măsurare pe scala ordinală, metodologia psihologică şi psihosociologică ar fi sărăcită de aportul statisticilor parametrice („tari"), pentru anumite măsurători se acceptă utilizarea scalei de interval – prima care permite calculul mediei şi a abaterii standard – însă, acest lucru se face printr-un important compromis statistico-matematic (Clocotici, şi alţii, 2000).

Variabile precum gradul didactic, gradul militar, nivelul de educație se găsesc la acest nivel de măsurare, deoarece pe lângă posibilitatea clasificării, a includerii unui individ într-o categorie, avem şi posibilitatea ordonării categoriilor, putând afirma că o categorie este superioară sau inferioară alteia.

Dacă luăm, de exemplu, gradele militare, putem spune că într-o unitate se află 30 de locotenenți, 10 maiori şi 60 de căpitani sau 15% locotenenți, 5% maiori şi 30% căpitani (ne aflăm la un nivel nominal de măsură) dar şi că maiorii sunt superiori în grad căpitanilor sau locotenenții sunt inferiori în grad căpitanilor (nivelul ordinal de măsură). Nu putem preciza însă cu cât şi nici de câte ori o categorie este superioară sau inferioară alteia. Nu putem afirma că doi locotenenți fac cât un căpitan. De asemenea, este absurd să vorbim despre medie. Care este media dintre un locotenent şi un căpitan??!!!

Proprietăți ale scalei:

Aşa cum am admis anterior, proprietățile specifice scalei ordinale le includ pe cele ale scalei nominale, astfel încât la cele ale scalei anterioare se adaugă (Vasilescu, 1992):

- Transformările permise la nivelul acestei scale sunt cele care nu afectează ordinea inițială. Astfel de operații sunt ridicarea la putere sau, inversa ei, extragerea de radicali. Din aceste operații, derivă:
- Operațiile statistice permise:
    - Frecvența sumelor în valori cumulate şi procentuale, valorile categoriale şi categoriale procentuale, valori categoriale centile, valori de împrăştiere;
    - Procedee bazate pe date categoriale: testul semnului, Mann-Whitney, Wilcoxon, Kolmogorov-Smirnov, analiza de varianță Kruskal şi Wallis;
    - Coeficienți de corelație de rang: ρ Spearman, τ Kendall.

I.2.2 Scale parametrice

Cuprind scalele de interval şi de raport. Începând de la acest nivel, putem vorbi de statistici parametrice sau metrice. Are sens de acum să calculăm

media, abaterea standard sau alți indicatori care fac parte din statisticile numite şi statistici „tari”.

I.2.2.1 Scale de interval (intervale egale)

Reprezintă al treilea nivel de măsurare şi derivă din scala ordinală, la care se adaugă proprietatea că intervalele dintre un nivel de valori şi altul sunt egale. Este un tip de măsurare în care distanțele dintre treptele scării sunt distanțe egale sub aspectul cantității caracteristicii de măsurat.

În psihologie practic nu întâlnim această scală decât în domeniul psihofiziologiei, în care înregistrăm timpi de reacție, număr de erori, forța unei reacții etc. Acest nivel de măsurare ne permite să răspundem la întrebarea „cu cât este mai mare” dar nu şi la întrebarea „de câte ori”, deoarece la nivelul acestei scale nu întâlnim un punct zero absolut, ci unul arbitrar ales.

Un exemplu clasic de scală de interval este scala de temperatură Celsius. După cum ştim, punctul de zero grade Celsius este un punct arbitrar ales, definit ca punctul de îngheț al apei la presiunea atmosferică de la nivelul mării. Intervalele acestei scale sunt egale, ceea ce ne permite să spunem că apa dintr-o găleată este mai caldă cu 10 grade Celsius decât apa dintr-o altă găleată, dar nu şi că apa este de două ori mai caldă, deoarece punctul zero nu este un zero absolut.

Prin analogie, ne putem referi şi la scorurile unui test de inteligență dacă admitem compromisul menționat în subcapitolul anterior. Putem spune că, între un subiect cu un IQ de 60 şi unul cu un IQ de 120 există o diferență de 60 de puncte, dar nu că cel cu un IQ de 120 este de două ori mai inteligent decât cel cu un IQ de 60. În definitiv, nu avem o inteligență zero. La o privire mai atentă observăm şi compromisul efectuat. Performanța la un test de inteligență depinde de mai mulți factori, nu numai de cei legați de inteligență. Astfel, intervin motivația, memoria, atenția, condițiile de examen etc. Este cert că cel de al doilea subiect este mai inteligent decât primul (ne situăm

acum la nivel ordinal), însă a afirma, chiar şi că este mai inteligent cu 60 de unităţi pe scala IQ, consider că este hazardat (la nivel de interval). În definitiv, la o reexaminare a celor doi subiecţi cu acelaşi test există o probabilitate foarte mare ca această distanţă, de 60 de puncte, să nu se păstreze. Iată o măsurare care, deşi la prima vedere pare la nivel de interval, totuşi, mai exact, o putem situa la nivel ordinal. Diferenţa dintre scala inteligenţei şi scala temperatorilor Celsius este totuşi foarte mare. Mai mult, intervalele pe o scală IQ sunt doar aparent egale. Diferenţa de 5 „puncte IQ" dintre un subiect cu un IQ de 60 şi unul cu un IQ de 65 are cu totul alt sens în comparaţie cu diferenţa dintre un subiect cu un IQ de 120 şi altul cu un IQ de 125 şi, de asemenea, se deosebeşte de diferenţa dintre doi subiecţi cu IQ 100 şi 105. Cele 5 „puncte IQ" nu sunt, în aceste cazuri, egal distanţate.

Proprietăţi ale scalei:

- Transformările permise la acest nivel sunt cele de tip liniar, caracteristice ecuaţiei $y=ax+b$ unde $a>0$ iar y este valoarea transformată, x – valoarea care urmează a fi transformată, a – constanta de extindere şi b – constanta de deplasare (Vasilescu, 1992).
    - Deplasări – prin mărirea punctului „zero" relativ la o altă valoare (modificarea constantei b);
    - Extinderi – amplificarea sau diminuarea intervalelor scalei cu acelaşi factor (modificarea constantei a), fără să deteriorăm rezultatul măsurătorilor (de exemplu transformarea din note „*z"* în stanine sau sten).
- Operaţii statistice permise:
    - Calculul mediei aritmetice, abaterii standard, boltirii şi simetriei;
    - Statistici parametrice: testul „t" Student, „F" – Fisher, analiza de varianţă;

- Toate tipurile de corelații: „r" Pearson, raport de corelație (R), coeficientul de regresie (b)

Anumite operații statistice, deşi presupun scala de interval, necesită şi îndeplinirea altor condiții şi anume respectarea unei legi de distribuție, în general legea distribuției normale (gaussiene). Prin urmare, în analiza datelor, pe lângă cerința nivelului de interval se impune şi analiza distribuției datelor înregistrate. Transformările permise la nivelul acestei scale au o importanță practică deosebită atunci când procedăm la normalizarea unei distribuții statistice de date.

I.2.2.2 Scala de raport (proporții)

Ultimul nivel de măsurare – şi cel mai precis – este reprezentat de scala de raport care are toate caracteristicile unei scale de interval la care se adaugă existența unui zero absolut. În sfârşit, această scală permite să răspundem la întrebarea „de câte ori", deoarece existența unui zero absolut face posibilă compararea proporțiilor.

De exemplu, dacă un subiect are greutatea de 60 de kilograme iar un altul 120 kilograme, putem spune că al doilea subiect este cu 60 de kilograme mai greu decât primul, dar şi că al doilea subiect este de două ori mai greu decât primul, ambele afirmații având sens.

Din nefericire, un asemenea nivel de precizie nu poate fi atins în ştiințele socio-umane, deoarece este imposibil să găsim o variabilă care să admită un zero absolut. Este absurd să vorbim de o inteligență zero, de o emotivitate zero, de depresie sau anxietate zero.

Proprietăți ale scalei (Vasilescu, 1992)**:**

- Transformările permise la acest nivel sunt cele de tip multiplicativ, caracteristice ecuației $y=ax$ unde $a>0$, iar y este valoarea transformată, x – valoarea ce va fi transformată iar a – constanta de extindere. Observăm dispariția constantei b, constanta de de-

plasare, deoarece punctul zero este un punct de referinţă absolut şi nu unul arbitrar ales. La fel ca şi în cazul scalei de interval, putem amplifica sau diminua intervalul (prin modificarea constantei *a*) însă nu putem opera deplasări, deoarece acest lucru ar determina scăderea preciziei către o scală de interval (am stabili un punct zero arbitrar şi nu absolut)

- Sunt permise toate operaţiile statistice, inclusiv calculul mediei geometrice şi al coeficientului de variaţie.

I.3 Prezentare generală SPSS for Windows[2]

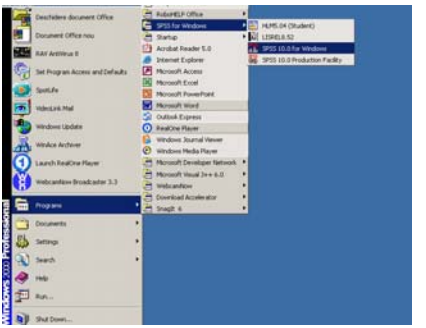


Figura 1.2 – Lansarea programului SPSS for Windows

SPSS sub Windows reprezintă un pachet de programe interactiv, de utilitate generală, destinat analizelor de date şi include multiple facilităţi şi tehnici de natură statistică. Pachetul de programe SPSS sub Windows reprezintă o aplicaţie care foloseşte întreaga funcţionalitate a sistemului de operare Windows, regăsind în structura sa utilizarea mausului, ferestrele redimensionabile şi scalabile, meniuri derulante, casete de dialog etc. Deşi majoritatea utilizatorilor vor fi complet satisfăcuţi utilizând doar interfaţa grafică pe care SPSS o pune la dispoziţie, totuşi pachetul de programe dispune şi de un puternic limbaj de comenzi prin care se pot realiza activităţi automate (scripturi) ce determină eficientizarea acţiunilor utilizate în

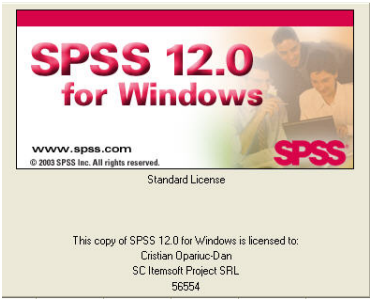


Figura 1.3 – Fereastra logo SPSS for Windows

[2] Pe parcursul acestei prezentări, unele imagini sunt din SPSS 12.0 for Windows. Acest lucru nu incomodează în utilizarea altor versiuni. Toate imaginile utilizate reprezintă marcă înregistrată SPSS Inc.

mod frecvent sau efectuarea unor acțiuni complexe, indisponibile în interfața grafică (SPSS, 2001).

Programul SPSS, la fel ca orice altă aplicație sub sistemul de operare Windows, indiferent de versiunea acestuia, la instalare își creează un grup de pictograme (iconuri) în secțiunea **Programs** a meniului **Start** din sistemul de operare Windows. Pentru a putea lansa în execuție aplicația, va trebui ca mai întâi să efectuați clic cu maus-ul pe butonul start al desktop-ului Windows, buton situat de obicei în partea din stânga jos a ecranului. Apoi mergeți la secțiunea **Programs** și se va deschide o nouă listă care conține toate programele instalate în calculatorul dumneavoastră (ei bine, cel puțin programele la care aveți acces dumneavoastră). În această listă căutați grupul de programe **SPSS for Windows**, grup în care veți găsi pictogramele de lansare ale aplicației SPSS. Probabil că ați identificat deja în acest grup de programe pictograma numită **SPSS for Windows**. Aceasta este imaginea pe care va trebui să faceți clic pentru a porni aplicația.[3]

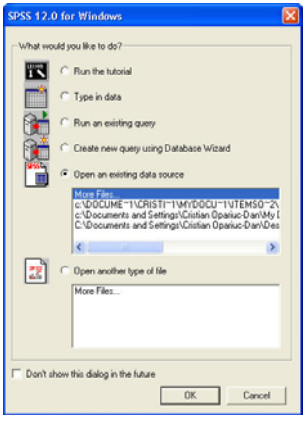


Figura 1.4 – Fereastra introductivă

SPSS este un program extrem de complex. Lansarea acestuia debutează cu prezentarea unei ferestre introductive, denumită și fereastră „logo". Ea cuprinde informații despre versiunea programului (în cazul nostru SPSS 12.0 Standard) deoarece există și versiuni server, mai complexe, versiuni demonstrative (ca aceea pe care o puteți descărca gratuit de la http://www.spss.com) sau „student" pentru învățare, etc. precum și despre posesorul licenței de utilizare a acestui produs informatic. Se cunoaște faptul că, în conformitate cu legea dreptu-

[3] Desigur, în condiția în care SPSS for Windows este deja instalat pe calculatorul dumneavoastră. Dacă nu aveți instalat SPSS for Windows, puteți descărca o versiune demonstrativă de 30 de zile de pe site-ul http://www.spss.com Asigurați-vă că aveți o conexiune suficient de rapidă și de stabilă la Internet deoarece fișierul este destul de mare (aproximativ 160 MB)

lui de autor şi protecţia creaţiilor intelectuale, un produs software poate fi utilizat doar în cazul în care se posedă o licenţă de utilizare pentru acesta, altfel riscându-se pedepse severe, contravenţionale sau de natură penală.

După câteva secunde, fereastra de prezentare se dezactivează iar aplicaţia prezintă o interfaţă prietenoasă prin care ne invită să trecem la treabă. Putem în această etapă să lansăm în execuţie un program tutorial care ne va forma abilităţile de bază în vederea utilizării acestui pachet informatic complex; de asemenea putem trece direct la introducerea datelor sau putem rula o cerere predefinită, se poate crea o nouă cerere de date sau se poate deschide o bază de date existentă, situaţie în care va trebui să precizăm locaţia acesteia. Nu în ultimul rând, putem deschide o bază de date într-un alt format, de exemplu Excel, Statistica sau SAS, iar SPSS va şti să convertească aceste date pentru a le face disponibile.

Există, de asemenea, în partea de jos a acestui formular introductiv, o casetă denumită **„Don't show this dialog in the future"** Dacă o veţi bifa, aveţi posibilitatea ca la viitoarea lansare a produsului să nu se mai afişeze acest formular iar programul va trece direct în fereastra principală SPSS.

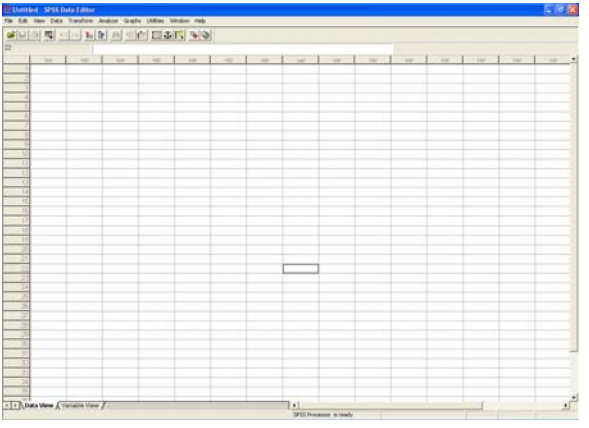


Figura 1.5 – Fereastra principală SPSS

Desigur, alegerea unei opţiuni dintre cele prezentate mai sus nu înseamnă şi executarea acesteia de către SPSS. Pentru a lansa în execuţie opţiunea aleasă, va trebui apăsat, obligatoriu, butonul **OK.** Apăsarea butonului **Cancel** indică

programului că trebuie să părăsească formularul introductiv şi să lanseze fereastra principală SPSS în modul de lucru configurat pentru o nouă bază de date.

În acest moment, nu ne interesează niciuna dintre opţiunile predefinite, astfel încât vom apăsa butonul **Cancel**. Iată că formularul prezentat mai sus a dispărut, făcând loc ferestrei principale SPSS. Să privim cu atenţie această nouă interfaţă cu utilizatorul.

Ca orice altă aplicaţie Windows, SPSS posedă o bară de titlu – bara albastră din partea de sus a ferestrei – în partea stângă a acesteia afişându-se mesajul **Untitled – SPSS Data Editor**. Acest mesaj ne informează că ne aflăm în fereastra de date (editare a datelor) din SPSS iar baza de date nu a fost încă salvată (în momentul salvării, textul „Untitled" este înlocuit cu numele fişierului salvat pe disc). În partea dreaptă a barei de titlu avem cele trei butoane clasice ale oricărei ferestre Windows: butonul de minimizare sau de transfer a aplicaţiei în bara de sarcini, butonul de maximizare – extindere a aplicaţiei pe întregul ecran sau de restaurare a aplicaţiei la dimensiunile iniţiale şi butonul de închidere a aplicaţiei, acel buton în formă de *„X"*. Sub bara de titlu se află meniurile, serii de etichete purtând fiecare un nume (în limba engleză), prin intermediul cărora putem comanda SPSS şi putem efectua operaţiile şi prelucrările de care acesta este capabil. Nu vom insista acum asupra explicării tuturor meniurilor, acest lucru îl vom face pe parcursul prezentului curs. Bara de meniuri este urmată de bara de instrumente – mici butoane cu aspectul unor pictograme care în general dublează anumite funcţii din meniuri, funcţii mai des folosite. SPSS prezintă o bară de instrumente dinamică, adică aspectul butoanelor şi desigur funcţionalitatea acestora se modifică relaţionat contextului în care ne aflăm. De exemplu, vom avea anumite butoane în modul de introducere a datelor şi alte butoane în modul de definire a variabilelor ori la afişarea rezultatelor analizelor de date.

După bara de instrumente, apare secţiunea de editare a datelor. Aceasta este formată dintr-o parte needitabilă, informativă, partea din stânga, în care se afişează în permanenţă poziţia celulei active (celulă pe care am selectat-o). Formatul acestui identificator este un număr urmat de două puncte şi apoi numele variabilei (spre exemplu, *8:time* înseamnă că suntem poziţionaţi pe rândul 8 – a opta înregistrare – a variabilei timp). Partea editabilă arată întotdeauna valoarea acestei variabile în poziţia specificată (în cazul nostru, la rândul 8 al variabilei timp se află valoarea 34).

Figura 1.6 – Secţiunea de editare a datelor

Cea mai mare parte a ferestrei principale SPSS este dedicată secţiunii de date. Observăm că datele, în SPSS, sunt organizate tabelar, pe linii şi coloane, analog altor aplicaţii mai cunoscute cum ar fi foile de calcul tabelar (Excel) sau bazele de date (FoxPro sau Access). Liniile (înregistrările) sunt identificate prin numere iar coloanele reprezintă variabilele. Într-o bază de date „goală" toate variabilele sunt denumite implicit *„var"* iar această denumire este scrisă cu culoarea gri, reprezentând faptul că acestea nu au fost încă definite. Navigarea prin baza de date se poate face cu cele două casete de derulare (numite şi lifturi sau controale de derulare) situate în partea dreaptă, respectiv în partea de jos a ferestrei de date, la fel ca în orice program Windows.

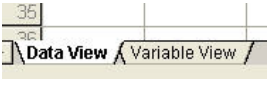


Figura 1.7 – Marcatorii de secţiune

Secţiunea de date conţine, în partea din stânga-jos, doi marcatori de secţiune foarte importanţi: **Data View** şi **Variable View.** Aceste elemente permit comutarea între modul de vizualizare a datelor dintr-o bază de date (secţiunea Data View) şi modul de definire al variabilelor (secţiunea Variable View).

În partea de jos a programului SPSS se află bara de status cu două secţiuni: secţiunea de informaţii, în partea stângă, unde se afişează scurte

instrucțiuni despre diferitele funcții ale meniurilor sau ale barei de instrumente şi secțiunea de stare a procesorului SPSS unde se afişează informații referitoare la rularea unor sarcini mari consumatoare de timp. În general, începerea unei prelucrări statistice se poate face doar în condițiile în care vedeți mesajul **SPSS Processor is ready** în această secțiune.

Părăsirea programului SPSS se poate face fie apăsând butonul de închidere al aplicației „X" din colțul din dreapta sus al ferestrei principale, fie folosind meniul **File** şi apoi **Exit**. Dacă ați făcut prelucrări de date sau modificări în baza de date, programul vă va solicita mai întâi salvarea documentului, ca măsură de protecție împotriva pierderii de informație. Despre salvare însă discutăm în subcapitolul următor.

I.3.1 Bazele de date – creare, salvare, deschidere

După ce ne-am familiarizat cu interfața programului SPSS, este momentul să vedem cum putem deschide o bază de date existentă şi care sunt elementele componente ale acesteia. SPSS vine însoțit de o serie de baze de date demonstrative în scopul facilitării procesului de învățare. Pe parcursul acestor exerciții vom folosi, din rațiuni de compatibilitate, aceste exemple de baze de date oferite de către producătorii programului SPSS.

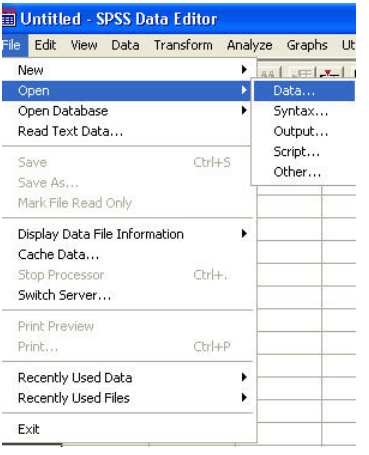


Figura 1.8 – Deschiderea unei baze de date

Deschiderea unei baze de date în scopul efectuării prelucrărilor statistice se poate face prin mai multe metode, rămânând la latitudinea şi la experiența dumneavoastră să alegeți metoda care vă convine mai mult.

Prima metodă este aceea prin care apelăm la meniul **File** (Fişier) şi apoi poziționăm cursorul maus-ului pe submeniul **Open** (Des-

chide). Apare o listă de unde putem deschide o bază de date (**Data),** o procedură de sintaxă (**Syntax**) o foaie de rezultate (**Output**), un set de comenzi (**Script**) sau un alt tip de fişier (**Other**). Pentru a deschide o bază de date, va trebui selectată opțiunea **Data.** Observați că fiecare dintre aceste opțiuni este urmată de puncte de suspensie (…). Întotdeauna când veți întâlni aceste puncte de suspensie alături de o comandă Windows (nu doar în SPSS ci în orice program Windows) va trebui să ştiți că acea comandă vă invită la dialog. Adică se va deschide o nouă fereastră de unde va trebui să alegeți ceva sau să furnizați anumite informații programului pentru a putea executa comanda respectivă.

O altă metodă prin care puteți deschide o bază de date este folosirea aşa numitelor „taste rapide". De obicei, meniurile sunt însoțite de combinații de taste prin care puteți efectua direct o anumită sarcină. În loc să apelăm la meniul File apoi Open şi apoi Data, putem pur şi simplu să ținem apăsată tasta **Ctrl** în timp ce apăsăm tasta **O** (de la Open). O astfel de combinație se notează prin **Ctrl+O**, notație pe care o vom folosi pe întreg parcursul acestui curs. Vom observa că această combinație are acelaşi efect ca şi parcurgerea primei metode, avantajul fiind acela că este mult mai rapidă.

A treia metodă apelează la numitele „taste fierbinți" (hot-keys). O „tastă fierbinte" este o tastă folosită în combinație cu tasta **Alt.** Dacă veți privi cu atenție meniurile, veți putea observa că unele litere sunt subliniate – de exemplu în meniul File este subliniată litera F. Ținând apăsată tasta **Alt** şi apăsând apoi tasta **F** vom obține deschiderea meniului File, ca şi când am efectua clic pe acesta. În mod analog, notarea acestei combinații se face folosind convenția **Alt+F**. Deci, pentru a deschide o bază de date putem folosi combinațiile **Alt+F** – deschiderea meniului File şi apoi **Alt+O** – deschiderea submeniului Open iar în final **Alt+A** – accesarea opțiunii Data. Este, dacă doriți, o navigare prin meniuri fără maus.

Ultima metodă se referă la folosirea barei de instrumente pentru a deschide o bază de date. Pur şi simplu apăsaţi pe primul buton de pe bara de instrumente, acela care are aspectul unui dosar deschis.

Indiferent de metoda folosită, efectul va fi acelaşi: deschiderea unei casete de dialog din care puteţi alege baza de date pe care o veţi încărca (vă mai amintiţi nota referitoare la punctele de suspensie?)

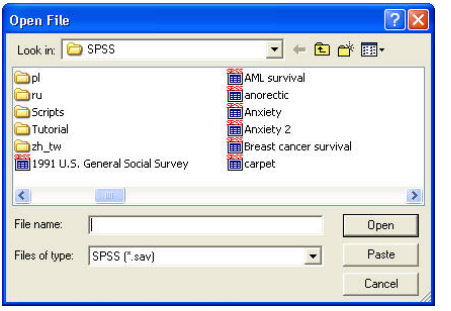


Figura 1.9 – Fereastra de selectare a fişierului în vederea deschiderii

Să analizăm puţin această nouă fereastră. În partea de sus avem o listă derulantă intitulată **Look in:** Apăsând săgeata orientată în jos partea dreaptă a acestei casete vom putea alege directorul în care este localizată baza noastră de date. Observăm că directorul curent este directorul aplicaţiei (SPSS) iar în secţiunea de sub această casetă este afişat conţinutul acestui dosar. Puteţi astfel selecta orice director în care aveţi fişiere recunoscute de SPSS pentru a le putea afişa şi/sau deschide. Alături de caseta **Look in** se află o serie de butoane care controlează navigarea prin structura de directoare de pe discul dumneavoastră. *Primul buton*, cel de forma unei săgeţi orientate în partea stângă, permite navigarea la ultima acţiune efectuată. Dacă, de exemplu, din directorul SPSS v-aţi poziţionat pe directorul Windows, apăsând acest buton veţi reveni din nou în directorul SPSS. La prima deschidere a acestei casete de dialog, butonul este inactiv deoarece nu aţi efectuat încă nici o acţiune care să fie înregistrată. *Al doilea buton*, de forma unui dosar cu o săgeată orientată în sus, permite navigarea ascendentă prin arborele de directoare. La o primă apăsare vă veţi deplasa în directorul Program Files, la următoarea vă veţi situa în directorul rădăcină C: şi aşa mai departe. *Cel de-al treilea buton* permite crearea unui nou director în directorul curent. Apăsând pe el veţi putea crea un nou director, numit New Folder, denumire pe care o veţi putea modifica după dorinţă. În sfârşit,

ultimul buton controlează modul în care se afişează pictogramele în caseta centrală. Puteţi alege vizualizarea pictogramelor în format mare, mic, sub formă de listă sau listă detaliată.

Fereastra centrală afişează, după cum am menţionat, structura de subdirectoare şi fişiere a directorului curent. Puteţi efectua dublu clic pe numele unui director pentru a afişa conţinutul acestuia sau pe numele unui fişier pentru a-l deschide în mod direct (fără a mai apăsa butonul Open). Dacă efectuaţi doar un simplu clic, numele fişierului selectat va fi afişat în caseta **File name** şi va trebui să folosiţi apoi butonul **Open** pentru a-l deschide. Caseta derulantă **Files of type** reprezintă un filtru pentru fişierele afişate. Dacă o vom deschide, efectuând clic pe săgeata orientată în jos din partea dreaptă a acesteia, vom putea alege ce anume tip de fişiere să se afişeze în secţiunea centrală. Observăm că, deşi directorul SPSS conţine mai multe fişiere, sunt afişate doar fişierele cu extensia *.sav*, adică fişierele ce conţin baze de date SPSS, deoarece în această casetă de filtrare s-a ales extensia *.sav*. Putem alege alte extensii de fişiere ce sunt afişate, de exemplu formatul Excel (*.xls*) sau dBase (*.dbf*) iar SPSS va afişa doar fişierele ce au această extensie.

Deschiderea propriu-zisă a unui fişier se face fie efectuând clic pe acesta şi apăsând butonul **Open** fie efectuând dublu clic pe numele fişierului. Spre exemplu, vom încărca baza de date „Cars" din directorul implicit SPSS, bază de date ce reprezintă un inventar al diferitelor tipuri de maşini şi a caracteristicilor acestora.

După deschidere observăm că s-a modificat aspectul ferestrei principale SPSS. În primul rând, în bara de titlu a apărut mesajul **Cars – SPSS Data Editor,** mesaj ce ne informează că am deschis baza de date „Cars" şi este pregătită pentru analiză. De asemenea, constatăm că structura acesteia este compusă din opt variabile, fiecare dintre ele reprezentând o anumită caracteristică a maşinilor. Ultima variabilă, variabila „Filter_$" este o variabilă internă folosită de programul SPSS pentru o filtrare anterioară a datelor, varia-

bilă pe care ne propunem să o ştergem. Cum? Efectuând clic pe numele acesteia. Observăm că întregul conținut al coloanei „Filter_$" s-a selectat. În continuare, apăsăm tasta **Delete**, situată pe tastatură, în partea dreaptă a grupului principal de taste, iar variabila a fost eliminată împreună cu toate datele acesteia. Felicitări. Iată şi prima dumneavoastră intervenție reală în SPSS. După cum bine ştim, orice primă intervenție într-un program înseamnă o „ştergere a ceva". Nici noi nu puteam face excepție!

Baza de date deschisă de către dumneavoastră a fost afişată în secțiunea **Data view** fapt care ne atrage atenția că în acest mod putem afişa şi urmări conținutul datelor din baza de date. Dacă vom efectua clic pe secțiunea **Variable view** vom intra în modul de definire al datelor. După cum ne spune şi numele secțiunii, aici putem vizualiza şi edita denumirea, tipul şi alte informații referitoare la variabile. Ca să folosim un limbaj mai tehnic, dacă în **Data view** ne aflăm în secțiunea datelor propriu zise, în **Variable view** ne aflăm în secțiunea meta-datelor, adică în secțiunea informațiilor despre date unde putem interveni asupra structurii lor interne. Vom clarifica imediat, prin câteva exemple, aceste aspecte.

Astfel, prima variabilă, denumită „mpg" este o variabilă numerică, ce acceptă maximum 4 caractere (adică cifre între 0 şi 9999) şi fără zecimale.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	mpg	Numeric	4	0	Miles per Gallo	None	None	8	Right	Scale
2	engine	Numeric	5	0	Engine Displac	None	None	8	Right	Scale
3	horse	Numeric	5	0	Horsepower	None	None	8	Right	Scale
4	weight	Numeric	4	0	Vehicle Weigh	None	None	8	Right	Scale
5	accel	Numeric	4	0	Time to Accele	None	None	8	Right	Scale
6	year	Numeric	2	0	Model Year (m	{0, 0 (Missing)}	0	8	Right	Ordinal
7	origin	Numeric	1	0	Country of Orig	{1, American}..	None	8	Right	Nominal
8	cylinder	Numeric	1	0	Number of Cyli	{3, 3 Cylinders	None	8	Right	Ordinal

Figura 1.10 – Secțiunea de definire a variabilelor

Numele „mpg" nu este însă un nume intuitiv şi nu ne putem da seama cu precizie la ce se referă această variabilă. Din păcate, doar versiunile de SPSS de

la 13.0 în sus acceptă nume de variabile cu mai mult de 8 caractere, însă nu se permite includerea spațiilor sau a altor semne de punctuație în denumirea acestora şi de aceea va trebui să fim foarte restrictivi în momentul în care le creăm. Pentru a înțelege mai uşor la ce anume se referă o anumită variabilă, avem posibilitatea de a o eticheta, adică de a include un text descriptiv asociat numelui ei. În cazul nostru, eticheta variabilei ne informează că „mpg" se referă la „Miles per Gallon" adică la consumul automobilului exprimat în distanța pe care acesta o poate parcurge cu un galon de combustibil.

Următoarea secțiune prezintă valorile pe care le putem defini pentru respectiva variabilă. Iată, în cazul variabilei „origin" care se referă la țara de producție a maşinilor observăm că s-a asociat valoarea 1 pentru maşinile de producție americană, valoarea 2 pentru cele de producție europeană şi valoarea 3 pentru cele japoneze.

În spațiul destinat definirii variabilelor avem de asemenea posibilitatea de a controla modul de tratare al cazurilor lipsă, alinierea datelor în cadrul celulelor şi definirea tipului de variabilă – nominală, ordinală şi scalară.

Efectuând clic pe aceste secțiuni ați observat că uneori apare în partea dreaptă a acestora fie un buton cu punctele de suspensie care va deschide o casetă de dialog unde se pot controla parametrii valorilor acceptate, fie un buton de incrementare/decrementare cu ajutorul căruia puteți mări respectiv micşora o anumită valoare, fie o casetă de selecție – săgeata orientată în jos – prin care puteți deschide o listă, de unde să selectați anumite valori.

Exercițiu practic:

În exercițiul ce urmează vom traduce pur şi simplu în româneşte structura acestei baze de date şi vom salva baza de date sub un nou nume, „Masini" (atenție, nu folosiți diacritica românească ş ci s. Deci nu vom salva cu „Maşini" ci „Masini") într-un director nou creat având calea „C:\Exercitii SPSS" (aceeaşi mențiune referitoare la diacritice).

În cazul primei variabile, numele va rămâne acelaşi, iar noi vom modifica doar descrierea. Vom da clic în secțiunea **Label** acolo unde se află afişat textul „Miles per Gallon” şi vom scrie textul „Mile per galon”.

A doua variabilă se numeşte „engine”. Vom schimba numele acesteia în „cilindree” (ne-am încadrat în 8 caractere), făcând clic în coloana **Name** şi scriind noul nume. Eticheta va deveni, pentru a doua variabilă, „Capacitate cilindrică”, pe care o veți modifica după exemplul primei variabile.

Procedăm la fel în cazul variabilelor 2, 3, 4, 5 şi 6, lăsând traducerea din engleză în română la latitudinea dumneavoastră. Atenție însă la acel maximum de 8 caractere când redenumiți o variabilă.

Variabila numărul 7 se referă la țara de origine a maşinilor. O vom redenumi „origine” şi vom modifica, desigur, eticheta. Există însă, la acest nivel, un element în plus şi anume coloana **Values**. Vă mai amintiți probabil de maşinile americane, europene şi japoneze. Dacă vom da clic pe celula corespunzătoare coloanei Values pentru această variabilă constatăm apariția unui mic buton cu puncte de suspensie în partea dreaptă a celulei. Apăsând acest buton se deschide o nouă fereastră, fereastra de configurare a valorilor asociate. Vorbeam anterior despre maşini de producție americană, europeană şi japoneză. În această casetă vedem lista acestor asocieri în partea de jos. În stânga listei există 3 butoane: **Add** cu ajutorul căruia putem crea o nouă asociere, **Change** prin care putem modifica o asociere selectată şi **Remove** care şterge, elimină asocierea selectată. Noi ne-am propus să traducem această structură de date în limba română. Vom efectua clic pe prima asociere 1 = „American”. Constatăm că, automat, în caseta **Value** a apărut cifra 1 iar în caseta **Value Label** textul „American”. Putem

Figura 1.11 – Etichetarea variabilelor de nivel nominal şi ordinal

deduce cu uşurinţă că valorii 1 i s-a asociat textul „American”(apropo, la ce nivel de măsurare ne aflăm aici şi ce fel de variabilă este aceasta?). De asemenea s-a activat butonul **Remove**, SPSS presupunând că dorim să eliminăm această asociere din moment ce am selectat-o. Pentru a modifica textul, efectuăm clic în interiorul casetei **Value Label** şi scriem în loc de „American”, „America” (mai bine spus, ştergem un „n”). De data aceasta de activează şi butonul **Change**, SPSS depistând modificarea efectuată şi presupunând că vom dori să o şi aplicăm efectiv. Dacă veţi apăsa acest buton (Change) veţi constata că eticheta s-a modificat din 1 = American în 1 = America. Felicitări. Aţi realizat prima traducere. Acum procedaţi la fel pentru „Europa” şi „Japonia”. Desigur, oricând puteţi include o nouă asociere. Tastaţi cifra 4 în caseta **Value** (deoarece ar urma 4 în ordine logică) şi textul „România” în caseta **Value Label** apoi apăsaţi butonul **Add** care aţi observat că a devenit activ. Astfel aţi asociat valoarea 4 maşinilor de producţie românească. În momentul în care aţi terminat de efectuat toate asocierile, apăsaţi butonul **OK** pentru a le memora şi a părăsi formularul. Reţineţi că apăsarea butonului **Cancel** duce la părăsirea formularului fără memorarea asocierilor.

Apare totuşi o întrebare. De ce să folosim aceste asocieri în loc de a scrie pur şi simplu în baza de date „America”, „Europa” şi „Japonia”? În primul rând pentru că este mai uşor. Imaginaţi-vă că aveţi de introdus un umăr de 3000 de date în baza dumneavoastră de date. În loc să scrieţi de fiecare dată un text format din 6-8 caractere, veţi tasta pur şi simplu 1 atunci când vă referiţi la America, 2 când vă referiţi la Europa şi 3 când vă referiţi la Japonia. SPSS va şti, de fiecare dată, în urma asocierilor realizate în etapa anterioară la ce anume se referă 1, 2 şi 3. În al doilea rând, deşi 1, 2 şi 3 sunt doar nişte etichete pentru America, Europa şi Japonia, se pretează mult mai bine prelucrărilor statistice ulterioare, fiind date cifrice.

Procedaţi la fel şi cu ultima variabilă, referitoare la numărul de cilindri ai motorului, pentru ca în final să obţineţi structura completă a bazei de date.

Prin acest exercițiu ne-am deprins cu unele elemente ale configurării structurii bazei de date. În continuare vom salva baza de date modificată sub denumirea de „Masini”. Directorul folosit va fi „C:\Exercitii SPSS” (atenție la diacritice).

Puteți alege meniul **File** şi opțiunea **Save** sau **Save as**, puteți apăsa combinația de taste **Ctrl+S** sau **Alt+F**, **Alt+S** sau puteți apăsa butonul de salvare de pe bara de instrumente (acela care are o dischetă drept pictogramă). Dar atenție! Diferența dintre opțiunea **Save** şi **Save as** este aceea că, folosind **Save** se vor salva modificările efectuate tot în baza de date inițială, în cazul nostru baza de date „Cars”. Noi ne-am propus să salvăm baza de date sub un alt nume şi într-un alt loc, deci nu vom putea folosi decât opțiunea **Save as** (Salvează ca). Se va deschide caseta de salvare a unei baze de date care are acelaşi aspect ca şi caseta de deschidere a unei baze de date, cu excepția butonului **Open**, în acest caz denumirea acestuia fiind **Save**. Folosind cunoştințele dobândite anterior, va trebui să vă poziționați cu ajutorul casetei **Save in** (echivalenta casetei **Look in** din fereastra de deschidere a bazei de date) în directorul rădăcină **C:** să creați un nou director folosind butonul de creare a unui director din acest formular pe care să-l denumiți **Exerciții SPSS**, să efectuați dublu clic pe acesta pentru a-l deschide şi să introduceți numele bazei de date **Masini** în caseta **File name**. Apăsați apoi butonul **Save** pentru a salva efectiv fişierul pe disc. Ați reuşit, nu-i aşa?

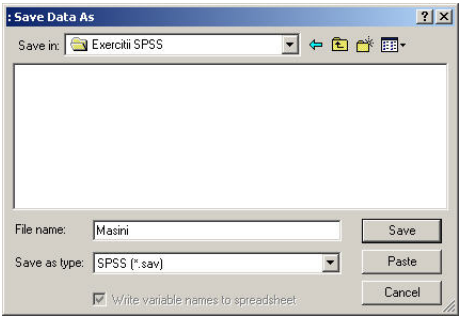


Figura 1.12 – Fereastra de salvare a unei baze de date

În urma acestor operațiuni, în bara de titlu SPSS va apărea mesajul **Masini – SPSS Data Editor** care ne informează că programul a salvat cu succes baza de date pe disc şi a activat-o.

Reamintim faptul că, la crearea unei noi baze de date, folosirea oricăror opţiuni de salvare (meniul File, butonul de pe bara de instrumente sau tastele rapide) duce la apariţia ferestrei **Save as** deoarece fiind o bază de date nouă, acesteia nu i-a fost atribuit încă un nume printr-o operaţie de salvare. Puteţi cu uşurinţă identifica dacă o bază de date are un nume (a fost salvată) sau nu, urmărind bara de titlu. Bazele de date noi sunt identificate prin mesajul **Untitled – SPSS Data Editor** în timp ce bazele de date salvate au afişat numele acestora în bara de titlu. Aţi observat că, la lansarea programului SPSS, acesta creează automat o bază de date nouă şi nedefinită şi aşteaptă din partea noastră definirea variabilelor şi salvarea acesteia. Odată salvată, orice modificări efectuăm în baza de date, la salvare ele vor fi scrise automat în fişierul al cărui nume este afişat în bara de titlu fără a mai apărea fereastra de salvare de mai sus. Singura modalitate prin care putem da un alt nume bazei de date sau prin care o putem salva într-un alt loc este folosirea opţiunii **Save as**.

Crearea unei noi baze de date se poate face folosind meniul **File**, submeniul **New** şi opţiunea **Data**, combinaţiile de taste **Ctrl+N** sau **Alt+F, Alt+N, Alt+A**. În oricare dintre aceste situaţii, SPSS va crea o bază de date goală şi nedefinită, la fel ca în cazul lansări programului.

Ca exerciţiu, folosind cunoştinţele dobândite, vă propunem crearea unei noi baze de date având următoarea structură şi salvarea acesteia sub numele de „IQ” în directorul C:\Exercitii SPSS”:

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	nume	String	20	0	Numele subiectilor	None	None	16	Right	Nominal
2	varsta	Numeric	3	0	Varsta subiectilor	None	None	9	Right	Scale
3	sexul	Numeric	1	0	Sexul subiectilor	{1, Masculin}...	None	8	Right	Nominal
4	scoala	Numeric	1	0	Nivel de scolarizare	{1, Primare}...	None	8	Right	Ordinal
5	iq	Numeric	3	0	Coeficient de inteligenta	None	None	8	Right	Ordinal

Variabila „nume” va fi de tip şir de caractere (string) şi va stoca numele subiecţilor, toate celelalte variabile fiind numerice. Există două definiţii

de asocieri şi anume variabila „sexul” (1-Masculin 2-Feminin) şi „scoala” (1-Primare 2-Gimnaziu 3-Liceul 4-Postliceala 5-Facultate)

După construirea structurii de date procedați la salvarea fişierului în locul precizat.

Folosind cunoştințele teoretice acumulate, precizați tipul fiecărei variabile creată în acest exercițiu.

Ne propunem în continuare aprofundarea definirii variabilelor în SPSS. Pentru aceasta ne vom folosi de baza de date „IQ” creată şi salvată în exercițiul anterior. În secțiunea **Data View** am remarcat deja prezența a zece caracteristici prin care se poate defini o variabilă în SPSS. Prima dintre acestea se referă la numele variabilei **(Name),** caracteristică ce nu presupune prea multe explicații. Numele variabilei reprezintă modul în care aceasta va fi apelată în toate prelucrările efectuate folosind baza de date. Nu putem însă să nu precizăm că SPSS impune o serie de restricții în denumirea variabilelor (versiunile mai mici de versiunea 13): în primul rând numele acestora să nu depăşească 8 caractere, variabilele cu nume mai lungi de 8 caractere nefiind acceptate. Apoi nu pot fi folosite spații sau semne de punctuație în denumirea variabilelor, putând fi însă folosită linia de subliniere pentru cazurile în care dorim să abreviem o variabilă formată din două cuvinte. De exemplu, dacă avem o variabilă ce se referă la anul naşterii, o vom putea abrevia denumindo *an_naste***.** Astfel, lungimea ei este de 8 caractere, fiind acceptată de SPSS iar cele două cuvinte componente le-am separat printr-o linie de subliniere, fapt de asemenea acceptat de SPSS. Nu contează dacă scriem numele variabilelor cu litere mari sau cu litere mici deoarece în final SPSS le converteşte automat în litere mici. Din punct de vedere al primei caracteristici, baza noastră de date IQ nu creează probleme, fiind definite cinci nume de variabile: *numele, varsta, sexul, scoala şi iq.*

A doua caracteristică se referă la tipul de date ce va fi stocat în aceste variabile (**Type**), caracteristică pe care nu am abordat-o în detaliu în timpul

prezentării noastre de până acum. Această secţiune conţine un buton cu trei puncte de suspensie pe care îl putem apăsa pentru a deschide fereastra de configurare a tipului de date. Aceasta conţine o serie de butoane de opţiuni dintre care putem selecta tipul de date pe care variabila noastră îl va stoca, după cum urmează:

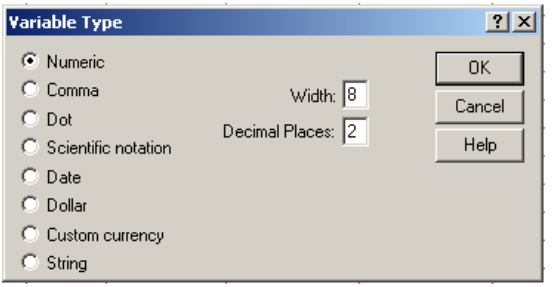


Figura 1.13 – Stabilirea tipului de date a unei variabile

Tipul de date „Numeric” – se referă la numere. Valorile acceptate vor fi deci numere, în format standard. Alegând acest tip de date, putem stabili şi dimensiunea acestora – valoarea maximă pe care SPSS o poate accepta – prin precizarea numărului de caractere în câmpul **Width** şi numărul de caractere după virgulă în câmpul **Decimal Places**. De exemplu, numărul maxim de caractere 3 şi 2 caractere zecimale înseamnă că acea variabilă poate stoca valori numerice de la 0 până la 999,99. Acest tip de date va scrie valorile numerice fără nici o separaţie între sute şi mii şi folosind ca separator zecimal separatorul configurat al sistemului. Un exemplu de număr scris în acest format ar fi 78457845.57

Tipul de date „Comma” – se referă de asemenea la numere şi pot fi definite la fel ca şi tipul anterior, numai că acestea vor fi afişate folosindu-se virgula ca separator între valoarea sutelor şi valoarea miilor şi punctul ca separator al valorilor zecimale. De exemplu, o valoare numerică în acest format ar arăta astfel: 9,999.89 Este practic formatul numeric clasic american.

Tipul de date „Dot” – identic cu tipul de date anterior, doar că se inversează rolul punctului şi al virgulei. Reluând acelaşi exemplu, numărul de mai sus va f scris astfel: 9.999,89 În acest caz recunoaştem formatul românesc.

Tipul de date „Scientific notation” – se referă la modul ştiinţific de afişare a cifrelor, folosind litera E pentru a reprezenta puterile lui 10. Modul de configurare al dimensiunii variabilei este acelaşi iar un exemplu de număr folosind această reprezentare poate fi 1.23E2 ceea ce semnifică de fapt 123.00

Tipul de date „Date” – se referă la date calendaristice. Alegând această opţiune va trebui să configuraţi, selectând din lista ce se va deschide, modul de afişare al datei calendaristice dintr-o mulţime de formate care vi se oferă.

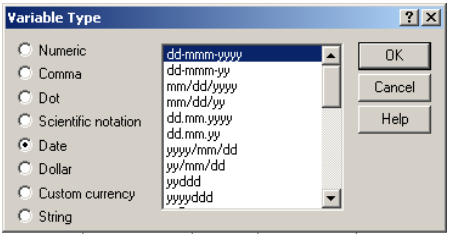


Figura 1.14 – Tipul de date calendaristic

Tipul de date „Dollar” – se referă la valori monetare avându-se în vedere moneda americană. Puteţi alege şi particulariza valoarea maximă pe care o poate accepta variabila din lista ce se afişează sau puteţi particulariza această valoare, la fel ca şi în cazul valorilor numerice.

Tipul de date „Custom currency” – permite definirea unor valori monetare particularizate, valabile pentru alte ţări în afara Statelor Unite. Se poate defini forma de prezentare a variabilei şi valoarea maximă acceptată, la fel ca şi la tipul anterior.

Tipul de date „String” – stochează doar şiruri de caractere, cu alte cuvinte texte, fără nici o posibilitate de prelucrare statistică. Poate fi definită şi aici dimensiunea maximă pe care o poate avea variabila din punctul de vedere al numărului de caractere acceptate. În cazul nostru, variabila „nume” este un exemplu excelent de variabilă de tip string.

Caracteristica a treia şi caracteristica a patra a unei variabile, respectiv **Width** şi **Decimals** se referă exact la valorile maxime pe care le pot accepta variabilele şi la numărul de zecimale, elemente pe care le-am discutat anterior

iar aceste două caracteristici nu fac decât să creeze o altă posibilitate de definire a variabilelor sub acest aspect.

În cazul exemplului nostru, constatăm că avem o singură variabilă de tip string, variabila „nume" care poate accepta maxim 20 de caractere, toate celelalte variabile fiind de tip numeric; variabilele „sexul" şi „şcoala" pot accepta doar o singură cifră fără zecimale iar variabilele „varsta" şi „iq" pot accepta 3 cifre fără zecimale. Această definiţie permite stocarea subiecţilor cu numele format din maxim 20 de litere, vârsta poate lua valori de până la 999, suficient deci pentru scopurile noastre, acelaşi lucru fiind valabil şi pentru coeficientul de inteligenţă. Sexul şi nivelul de şcolarizare acceptă doar o singură cifră, deoarece am convenit să asociem valoarea 1 cu sexul masculin şi 2 cu sexul feminin, în mod analog procedând şi cu nivelul de şcolarizare. Datele au fost definite ca numerice pentru a crea posibilitatea interpretării lor ulterioare, cunoscându-se faptul că datele de tip string nu pot fi interpretate statistic.

Următoarea caracteristică, caracteristica **Label,** permite includerea unei denumiri mai intuitive a variabilei, prin care să poată fi mai uşor identificată şi recunoscută. Astfel, variabilei „nume" îi putem da eticheta „Numele subiecţilor", variabila „iq" o putem eticheta ca şi „Coeficientul de inteligenţă" şi aşa mai departe. Putem include orice etichetă descriptivă dorim pentru numele unei variabile, cu condiţia ca lungimea acesteia să nu depăşească 255 de caractere. Deci atenţie la scrierea romanelor în descrierea variabilelor!!! Această caracteristică mai are însă o funcţie importantă; SPSS include eticheta variabilelor în rapoartele finale rezultate în urma prelucrării datelor. Dacă creaţi un tabel de frecvenţă, SPSS nu va afişa numai numele variabilei, de exemplu „iq" ci va include şi descrierea acesteia, „Coeficientul de inteligenţă", făcând astfel mult mai estetice şi mai inteligibile foile de raport.

Caracteristica **Values** am discutat-o pe larg în secţiunile anterioare şi nu vom mai reveni asupra ei. Menţionăm doar că ea permite clasificări (deci

variabile categoriale) şi asocieri de valori. Spre exemplu, putem asocia valoarea 1 sexului masculin şi valoarea 2 sexului feminin pentru a putea introduce mai uşor datele şi pentru a putea efectua prelucrări statistice la nivel de variabile nominale.

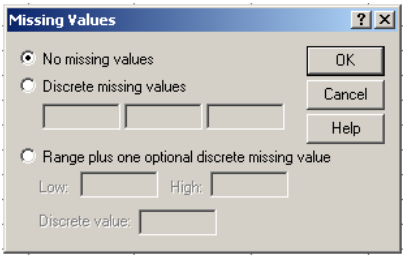


Figura 1.15 – Tratarea valorilor lipsă

Caracteristica **Missing** permite definirea modului în care vor fi tratate valorile lipsă din cadrul unei variabile. Uneori este important să cunoaştem de ce lipsesc valori dintr-o variabilă. Datele lipsă se pot datora, spre exemplu, faptului că subiectul refuză să răspundă la o anumită întrebare sau faptului că întrebarea nu este adresată acelui subiect. Iată două situații care ar necesita un tratament separat al valorilor lipsă. Datele pe care le veți trata în mod explicit ca date lipsă pot fi supuse apoi unei analize statistice separate, fiind inițial excluse din cele mai multe calcule. Un alt exemplu, frecvent întâlnit în chestionarele sociologice se referă la codificarea răspunsurilor de tip „Nu ştiu/Nu răspund". Majoritatea cercetătorilor codează aceste variante cu 9 sau 99. Evident, variantele incerte nu vor face obiectul prelucrării inițiale a datelor. Folosind această secțiune se pot include aceste valori (9 sau 99) ca valori lipsă, SPSS eliminându-le din analiză. Definirea valorilor lipsă poate accepta trei situații. Situația **„No missing values"**, când comunicăm programului că nu dorim un tratament special pentru valorile lipsă, acestea fiind tratate ca atare în cursul fiecărei prelucrări de date pe care o realizăm, situația **„Discrete missing values"** când putem defini până la trei valori ce vor fi tratate ca valori lipsă în cursul prelucrărilor statistice ulterioare (de exemplu, la un chestionar cu 3 variante de răspuns, dorim ca doar răspunsurile 1 şi 3 să fie luate în calcul, răspunsul 2 fiind tratat ca non-răspuns. Pentru aceasta vom include cifra 2 în prima dintre cele trei casete ale acestei secțiuni.) şi situația **„Range plus one optional discrete missing value"** în care putem stabili un interval

pentru care valorile vor fi tratate ca valori lipsă (spre exemplu, coeficienţii de inteligenţă cuprinşi între 10 şi 75) în plus existând posibilitatea de a include şi o valoare particulară (discretă) ca şi în situaţia anterioară. În exemplul nostru de bază de date, nu avem nici un motiv de a trata în vreun fel valorile lipsă, deci nu vom configura această caracteristică.

Caracteristica **Columns** permite specificarea lăţimii coloanei în care se vor afişa valorilor variabilelor în modul Data View. Implicit, coloanele au o lăţime egală cu 8 caractere. Putem include aici o altă valoare sau putem redimensiona lăţimea coloanelor în modul Data View prin „tragerea" cu ajutorul maus-ului a liniei de separaţie dintre două coloane. Introducerea diferitelor valori pentru această caracteristică permite ajustarea şi redimensionarea precisă a lăţimii coloanelor. În cazul nostru, pentru variabila „nume", dimensiunea implicită de 8 caractere va fi insuficientă pentru prezentarea unui subiect cu numele „Popescu Valentin", afişându-se doar primele 8 caractere şi anume „Popescu " Pentru a vedea întregul nume, fie introduceţi valoarea 16 pentru caracteristica Columns a variabilei „nume", fie trageţi cu ajutorul maus-ului linia de demarcaţie dintre coloana „nume" şi coloana „varsta" pentru a mări lăţimea coloanei de afişare a numelui subiecţilor. Această caracteristică, ca şi următoarea, nu influenţează în nici un fel modul de prelucrare al datelor ci doar aspectul tabelului de date din Data View.

Caracteristica **Align** controlează modul în care sunt aliniate valorile variabilelor în tabelul de date. Puteţi opta pentru o aliniere la stânga, centrare sau aliniere la dreapta a valorii variabilelor în cadrul coloanelor din Data View, fără a influenţa în vreun fel modul de prelucrare al variabilelor. Ca regulă generală, variabilele de tip string se aliniază la stânga iar variabilele de tip numeric sau dată calendaristică se aliniază la dreapta pentru o mai bună lizibilitate în cadrul tabelelor de date.

Ultima caracteristică, **Measure,** este în acelaşi timp şi una foarte importantă, permiţând alegerea nivelului de măsurare al variabilei. Se poate

opta pentru două sau trei niveluri de măsurare în funcție de tipul de variabilă. Astfel, pentru variabila „nume", care este o variabilă de tip string, sunt disponibile doar nivelurile nominal şi ordinal de măsurare. Stocarea numelui şi prenumelui are o valoare pur informativă şi nu permite nici măcar o clasificare a acestora. Nivelul de măsură va fi deci nivelul nominal. Variabila „varsta" este o variabilă numerică care stochează vârsta subiecților. Această variabilă are intervale egale, zero absolut şi permite rapoarte în cadrul scalei, deci poate fi considerată la un nivel de măsurare scalar[4]. Variabila „sexul" nu este altceva decât tot o variabilă nominală care nu permite decât o clasificare a persoanelor în funcție de genul lor biologic. Variabila „scoala" poate fi privită ca o variabilă nominală dacă intenționăm doar să clasificăm subiecții în funcție de studiile absolvite sau se poate accepta un nivel ordinal de măsurare la nivelul acesteia dacă ierarhizăm subiecții în funcție de gradul lor de şcolarizare. Vom lua în considerare a doua situație şi o vom accepta ca variabilă ordinală. În sfârşit, variabila „iq" este fără discuție o variabilă de tip ordinal.

Iată că, parcurgând aceste etape, am reuşit în cele din urmă să construim structura unei baze de date simple în acord cu principiile şi domeniile de definiție ale variabilelor şi cu nivelul lor de măsurare.

Un aspect important pe care merită să-l menționăm se referă la procesul de codare al variabilelor (proces pe care l-am întâlnit la codarea variabilelor sex şi nivel de şcolarizare). În general, cercetătorii cu experiență cunosc intuitiv modalitatea de codare optimă a unei variabile. Pentru a veni în sprijinul debutanților, Newton şi Rudestam (1999) propune, în lucrarea „*Your statistical consultant*" un set de reguli folosite în acest proces (Field, 2005).

- **Toate datele trebuie codate numeric.** Pentru a coda, de exemplu, sexul, am putea folosi şi „M" pentru masculin respectiv „F" pen-

[4] Ca observație, în SPSS nivelurile de măsură de interval şi de raport au fost reunite sub denumirea generică de nivel de măsură scalar. Proiectanții acestei aplicații au renunțat la distincția separată a celor două reunindu-le sub denumirea de nivel scalar.

tru feminin. Un asemenea proces de codare ar fi complet eronat, deoarece ar limita radical posibilităţile de prelucrare statistică a datelor. Recomandarea este aceea că, atunci când suntem în astfel de situaţii este de preferat să folosim, de exemplu, 1 pentru Masculin şi 2 pentru Feminin. Aceste cifre nu au decât o valoare de etichetă şi nu reprezintă un proces de ierarhizare (nu spunem că masculinul este „mai jos" decât femininul deoarece are valoarea 1). La fel de bine putem să scriem 2 masculin şi 1 feminin sau 5 masculin şi 7 feminin. Acest proces de codare poartă numele de **codare dummy.** Tot ceea ce trebuie să faceţi este să reţineţi codurile atribuite.

- **Fiecare variabilă trebuie să ocupe o coloană distinctă în baza de date şi fiecare subiect va ocupa o singură înregistrare.** Este foarte important ca datele unui subiect să ocupe un singur rând în baza de date SPSS, iar fiecare coloană să se refere doar la o singură variabilă. În cazul în care o variabilă este măsurată de mai multe ori (tipic pentru experimente de tip testare iniţială – testare finală), se vor aloca două coloane pentru acea variabilă, una care să se refere la testarea iniţială şi una vizând testarea finală. De exemplu, dacă avem un experiment prin care în urma unei tehnici oarecare presupunem că îmbunătăţim memoria, va trebui să testăm memoria subiecţilor înainte de aplicarea tehnicii, să aplicăm tehnica şi apoi să testăm din nou memoria subiecţilor, pentru a vedea dacă au apărut modificări. În acest caz, putem defini cele două variabile prin *mem_ini* ce reprezintă testul iniţial de memorie şi *mem_fin* care se referă la testul final. Astfel, nu încălcăm această regulă, iar datele vor fi consistente.
- **Toate codurile pentru o variabilă trebuie să fie mutual exclusive.** Cu alte cuvinte, nu trebuie să existe decât un singur număr ce

poate fi stocat într-o variabilă. În anumite chestionare pot exista itemi care permit alegeri multiple (de exemplu la întrebarea: *„Bifați din următoarea listă partidele cu care în mod sigur nu veți vota la următoarele alegeri:”* pot fi variantele de răspuns *„PSD”, „PD”, UDMR”, „PNG”* etc. În acest caz, subiectul poate alege un singur partid, două partide sau le poate alege pe toate). O asemenea situație impune crearea unui număr de variabile egal cu variantele de răspuns (de exemplu *nu_PDS, nu_PD, nu_UDMR, nu_PNG*) cu categorii dihotomice de genul *0 – Neselectat, 1 – Selectat* care vor fi analizate ulterior. De obicei, asemenea itemi au şi o categorie de genul *Altul. Care_____* prin intermediul căreia subiectul poate să-şi exprime liber opțiunea. În mod normal, această categorie se analizează prin alte metode şi nu folosind SPSS, deoarece variabilitatea răspunsurilor conferă o valoare statistică foarte redusă.

- **Fiecare variabilă trebuie codată astfel încât să se obțină maximum de informații.** Categoriile sau valorile variabilei nu trebuie sintetizate în prima variantă, ci trebuie incluse exact aşa cum apar pe formularele de colectare a datelor. Sintetizarea valorilor şi cumularea categoriilor poate fi făcută mai târziu, utilizând funcțiile specifice SPSS. De exemplu, se preferă introducerea exactă a valorii pentru vârsta subiectului, în ani, şi nu încadrarea subiectului în categorii de vârstă (sub 14 ani, între 14 şi 20 de ani, peste 20 de ani etc.). Inițial se va proiecta o variabilă, *varsta,* care va stoca vârsta exactă a subiecților. Ulterior, dacă se va dori transformarea acestei variabile scalare într-o variabilă discretă, ordinală, numită *grup_varsta*, se pot folosi tehnici specifice de conversie. Ca idee generală, întotdeauna când este posibil, se vor folosi variabile continui la un nivel de măsură cât mai aproape de scala de raport. Este uşor să transformăm o variabilă continuă într-una discretă, însă

imposibil să trecem de la o variabilă discretă la una continuă. După cum ştiți deja, valoarea informativă cea mai ridicată o au totuşi variabilele continui, astfel încât profitați de ele.

- **Fiecare subiect trebuie să aibă o valoare pentru fiecare variabilă.** Uneori nu este posibil să includem valori în fiecare coloană, fie deoarece subiectul nu a răspuns, fie pentru că itemul respectiv nu i se adresează. În acest caz se recomandă să lăsați acea celulă liberă, deoarece SPSS este astfel proiectat încât să trateze celulele goale ca şi cazuri lipsă. Introducerea altor valori pentru cazurile lipsă poate crea confuzii. Respectați întotdeauna regula „ceea ce e cel mai simplu este şi cel mai bun".

- **Regulile de codare se vor aplica tuturor subiecților.** Acest lucru înseamnă că nu puteți aplica reguli valabile doar pentru anumiți participanți. De exemplu, dacă doriți ca anumite valori să fie tratate ca valori lipsă pentru un anumit subiect, nu puteți face acest lucru. Valorile lipsă vor fi tratate similar pentru toți subiecții.

- **Folosiți coduri în concordanță cu sensul scalei.** Unele chestionare presupun tipuri de scale ordinale (de exemplu 1 – niciodată, 2 – rareori, 3 – uneori, 4 – deseori, 5 – întotdeauna). Deşi scalele pot fi codate şi invers (1 – întotdeauna, 2 – deseori, 3 – uneori, 4 – rareori, 5 – niciodată) se recomandă păstrarea sensului natural al scalei pentru evitarea problemelor ulterioare referitoare la interpretare. Mai uşor asociem niciodată cu 1 decât niciodată cu 5. Oricum, chiar dacă ulterior există suficiente posibilități de recodare a valorilor astfel încât analiza să fie consistentă, proiectarea adecvată a cercetării şi codării vă poate scuti ulterior de bătăi de cap şi muncă inutilă.

După proiectarea bazei de date se recomandă listarea structurii acesteia (meta-informațiilor) în vederea consultării ulterioare. Dacă nu veți lucra

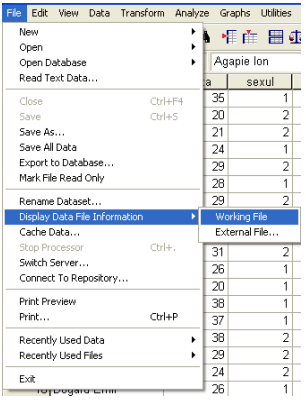


Figura 1.16 – Afişarea informaţiilor despre baza de date

câteva luni cu baza de date, probabil că nu veţi reţine la ce se referă mem_ini sau mpg şi nici de ce tip este. SPSS prezintă o modalitate uşoară de a afişa structura bazei de date cu informaţii complete. Nu trebuie decât să accesaţi meniul **File** şi apoi opţiunea **Display Data File Information.** Se va deschide o nouă secţiune care vă invită să alegeţi între a afişa informaţiile bazei de date curente (**Working File**) sau a unei alte baze de date de pe disc (**External File…**). Noi suntem interesaţi de afişarea structurii bazei noastre de date (IQ.sav) şi vom alege aşadar prima opţiune.

În fereastra de rezultate (**Output**) vor apărea o serie de tabele corespunzătoare definiţiilor variabilelor din baza de date, astfel:

Tabelul 1.1 – Tabelul de structură a bazei de date

Variable Information

Variable	Position	Label	Measurement Level	Column Width	Alignment	Print Format	Write Format
nume	1	Numele subiectului	Nominal	14	Left	A20	A20
varsta	2	Varsta subiectilor	Scale	8	Right	F2	F2
sexul	3	Genul biologic al subiectilor	Scale	8	Right	F1	F1
iq	4	Coeficientul de inteligenta	Scale	8	Right	F3	F3

Variables in the working file

Variable Values

Value		Label
sexul	1	Masculin
	2	Feminin

Tabelul **Variable Information** conţine informaţii despre variabile. Prima coloană conţine numele variabilelor, urmează apoi poziţia acestora în structura de date, eticheta, nivelul de măsură, lăţimea coloanei şi alinierea.

Ultimele două coloane se referă la tipul variabilei. Astfel, prima variabilă poate stoca 20 de caractere, a doua 2 caractere, a treia un caracter şi a patra 3 caractere.

Următorul tabel conţine variabilele codate. În cazul nostru singura variabilă codată este sexul şi conţine valorile 1 – Masculin şi 2 – Feminin.

Vă recomand ca după ce proiectaţi baza de date să imprimaţi structura acesteia şi să o stocaţi pentru a o putea consulta ori ce câte ori aveţi nevoie.

În concluzie:

- *Măsurarea în ştiinţele socio-umane are un caracter subiectiv, deoarece nu putem observa direct indicatorii funcţiilor psihice;*
- *Măsurătorile sunt aproximări ale realităţii deoarece sunt supuse în permanenţă surselor de eroare iar realitatea socială are un caracter fluctuant;*
- *După S. S. Stevens, „în sensul său larg, a măsura înseamnă a atribui numere obiectelor sau evenimentelor potrivit unor reguli."*
- *Regula de atribuire trebuie să fie:*
    - *Suficient de clară pentru a permite atribuirea univocă, la un moment dat, a unui număr unui obiect şi numai a unui singur număr;*
    - *Suficient de simplă pentru a fi utilizată;*
- *O variabilă este un nume pentru un element a cărui caracteristică este aceea că variază, îşi modifică valorile;*
- *O variabilă este aleatorie şi poate lua valori numerice (alea-numerică) sau nenumerice;*
- *O variabilă se caracterizează printr-un domeniu de definiţie, o mulţime de valori pe care aceasta le poate lua. În funcţie de acest domeniu de definiţie, variabilele pot fi discrete sau continui;*
    - *O variabilă discretă nu permite existenţa unor valori intermediare între valorile pe care le poate lua variabila. Ea este definită pe o mulţime numărabilă de valori în baza unei funcţii de probabilitate;*
    - *O variabilă continuă permite existenţa unor valori intermediare între două valori ale acesteia iar acest lucru poate continua la infinit. Ea este definită*

pe un interval de valori, mulțimea fiind infinită iar legea de distribuție este o densitate de probabilitate;

- *Scalele de măsurare se grupează în scale neparametrice (nominală şi ordinală) şi scale parametrice (de interval şi de raport);*
- *Termenul de cuantificare are sens începând cu scala ordinală;*
- *Scala nominală prezintă doar posibilitatea unor clasificări, grupări ale datelor în categorii;*
- *Scala ordinală permite, în plus, ierarhizarea datelor în funcție de apartenența lor la o categorie. Strict, aceasta este scala utilizată în psihologie;*
- *Scala de interval este o scală ordinală în care intervalele sunt egale. Ea permite să răspundem „cu cât o cantitate este mai mare decât o alta" însă nu şi „de câte ori";*
- *Scala de raport este o scală de interval la care există un zero absolut. Ea permite să răspundem la întrebarea „de câte ori o cantitate este mai mare decât cealaltă"*

Întrebări şi exerciții:

- *În bara de titlu a aplicației SPSS se află afişat mesajul „Untitled – SPSS Data Editor". Ce se va întâmpla când vom apăsa butonul de salvare a datelor de pe bara de instrumente după ce am proiectat structura unei baze de date?*
- *Am primit de la un alt colaborator o bază de date cuprinzând rezultatele unor probe psihologice, realizată însă în Excel. Ştiu că SPSS poate prelucra şi foi de calcul Excel, deci introduc discheta în calculator, folosesc procedura de deschidere a unei baze de date din SPSS, mă poziționez pe unitatea de dischetă (discul A:) însă nu reuşesc să văd fişierul Excel. Unde am greşit?*
- *Am creat baza de date IQ. De ce atunci când doresc să introduc numele subiectului Alexandrescu Delia Ioana Alexandra, la un moment dat SPSS nu mă mai lasă să tastez?*
- *Precizați de ce tip sunt următoarele variabile şi la ce nivel de măsurare se află: „viteza vântului", „gradele didactice dintr-o universitate", „anxietatea totală", „fabricanții de maşini", „echipele de fotbal", „numărul de tablouri dintr-o galerie de artă"*
- *Deşi subiectul „Popescu Vasile" are mai puțin de 20 de caractere, totuşi în momentul în care îl introduc în SPSS se afişează doar „Popes". Cum pot să fac pentru a vedea întregul nume al subiectului meu?*
- *Am peste 400 de subiecți în baza mea de date IQ şi doresc să efectuez o analiză în care toți coeficienții de inteligență mai mici de 70 să fie tratați ca şi cazuri lipsă. Cum procedez?*

II. ORGANIZAREA DATELOR

În acest capitol se va discuta despre:

- Colecţii de date brute;
- Amplitudinea de variaţie;
- Diagrame de date;
- Clase şi intervale de grupare;
- Efective, frecvenţe absolute şi relative;
- Reprezentări grafice.

După parcurgerea capitolului, cursanţii vor fi capabili să:

- Înţeleagă elementele de bază ale unei distribuţii statistice;
- Calculeze numărul de clase şi intervalele de grupare;
- Grupeze elementele în clase;
- Construiască o distribuţie statistică;
- Reprezinte grafic o distribuţie statistică.

Orice experiment, evaluare psihologică, observaţie sau convorbire porneşte de la o întrebare care-şi propune să rezolve o anumită problemă. De cele mai multe ori, problema în ştiinţele socio-umane nu transpare imediat într-o formă cuantificabilă. Avem nevoie de o serie de etape intermediare pentru a stabili ce dimensiuni măsurabile sunt în relaţie cu problema sau tema studiată. Procesul prin care transformăm o dimensiune generală, necuantificabilă, în indicatori măsurabili şi apoi construim itemi pentru acei indicatori poartă numele de **operaţionalizare a conceptelor**. Dacă, de exemplu, dorim să studiem nivelul de trai a populaţiei dintr-un oraş, o asemenea abordare este mult prea generală. Ce înseamnă în definitiv nivel de trai? Dimensiunea nu poate fi cunoscută direct, astfel încât apelăm la anumiţi indicatori care au legătură cu nivelul de trai (bugetul alocat pentru achiziţionarea de produse electronice, bugetul pentru concediu etc.) Urmează apoi să construim itemi, întrebări referitoare la indicatorii stabiliţi („Unde vă petreceţi concediile?”, „Ce produse electronice aţi cumpărat în ultimele trei luni?” şi aşa mai depar-

te). Studiind răspunsurile la aceste întrebări putem avea o imagine generală asupra nivelului de trai. Aceşti indicatori nu reprezintă altceva decât variabile rezultate în urma operaționalizării conceptului general de nivel de trai şi care pot fi populate cu datele rezultate din răspunsurile subiecților la întrebările stabilite.

În urma procesului de operaționalizare a conceptelor rezultă una sau mai multe variabile, reprezentative pentru studiul propus, variabile ce vor trebui „populate" cu date în vederea verificării ipotezelor şi a găsirii răspunsului la întrebarea formulată inițial. Datele cu care urmează a fi „populată" respectiva variabilă pot proveni din (Radu, şi alţii, 1993):

- Lectura unor aparate;
- Măsurători antropometrice şi/sau bio-fiziologice;
- Notele brute (răspunsurile subiecților) la anumite teste psihologice sau chestionare;
- Protocolul de observație sau interviu;
- Analiza de conținut sau analiza produselor activități, etc.

În practică putem fi interesați despre modul în care se grupează datele în jurul unor valori, de variabilitatea acestora sau de stabilirea poziției relative a unui individ sau a unui rezultat în cadrul unui grup mai mare de indivizi sau rezultate.

Fie că desfăşurăm o evaluare psihologică, un studiu sau o cercetare experimentală, fie că inițiem un proces de colectare a datelor prin observație, rezultatul constă într-o **colecție (set) de date brute**. La acest nivel nu avem nici o modalitate prin care putem trage concluzii asupra semnificației datelor pe care le avem la dispoziție. Putem doar constata că unele valori se repetă şi … cam atât. De aceea, prima etapă în inițierea unei proceduri statistice de analiză a datelor constă în clasificarea, ordonarea, condensarea acestora în

vederea conturării unei imagini cât mai precise. Valorile măsurate trebuie aşezate într-o anumită ordine (crescătoare sau descrescătoare) care să permită şi o eventuală reprezentare grafică a acestora.

Exemplu: Presupunem că măsurăm înălţimea unor subiecţi în centimetri şi obţinem următoarele rezultate: 182, 175, 174, 189, 177, 177, 180, 173, 188, 171, 184, 192, 180, 172, 177, 193, 184, 192, 172, 177, 174, 171, 172, 179, 185, 179, 183, 172, 179, 188.

***Întrebare:** Ce tip de variabilă este variabila menţionată mai sus şi la ce nivel de măsurare se află?*

Dacă privim şirul de date de mai sus, observăm că nu putem extrage nici o informaţie utilă. Unele valori se repetă, e adevărat, însă nu putem spune în lipsa unei organizări nici care este cel mai înalt subiect, nici care este cel mai scund, nici ce înălţime au cei mai mulţi subiecţi. Acest lucru va deveni posibil abia în urma unui proces de organizare a acestor date.

La volume mici de date acest proces nu pune probleme deosebite. Operaţiunea devine însă greoaie şi mare consumatoare de timp pentru volume mari de date şi de aceea se recomandă utilizarea unui program computer care poate efectua aproape instantaneu aceste operaţii.

II.1 Sistematizarea datelor

Primul pas în sistematizarea datelor îl reprezintă stabilirea valorilor extreme. Prin inspecţie vizuală, verificăm cea mai mică valoare din şirul de date (X_{min}) şi cea mai mare valoare din şirul de date (X_{max}) pentru a putea calcula poate cel mai simplu indicator statistic şi anume amplitudinea de variaţie. Numim **amplitudine de variaţie (Range)** diferenţa dintre valoarea maximă şi valoarea minimă a distribuţiei de date. $(A = X_{max} - X_{min})$. În exemplul nostru, subiectul cel mai scund (cu înălţimea cea mai mică) are 171 centimetri iar subiectul cel mai înalt are 193 centimetri. Prin urmare, $X_{min} = 171$

centimetri, X_{max} = 193 centimetri iar amplitudinea de variație devine *A = 193 – 171 = 22 centimetri.*

Dacă amplitudinea de variație este mică comparativ cu numărul de valori din şirul de date, atunci şirul de date poate fi uşor reprezentat fără necesitatea grupării valorilor. În cazul nostru, putem sistematiza datele scriind într-o coloană valorile în ordine crescătoare şi notând în dreptul fiecăreia, prin linii verticale, de câte ori se repetă în şirul inițial. Dacă reconsiderăm exemplul anterior, obținem datele grupate în tabelul alăturat:

Tabel 2.1 – Diagrama în linii şi efectivul

Valoare	Diagrama în linii	Efectiv
171	||	**2**
172	||||	**4**
173	|	**1**
174	||	**2**
175	|	**1**
176		**0**
177	||||	**4**
178		**0**
179	|||	**3**
180	||	**2**
181		**0**
182	|	**1**
183	|	**1**
184	||	**2**
185	|	**1**
186		**0**
187		**0**
188	||	**2**
189	|	**1**
190		**0**
191		**0**
192	||	**2**
193	|	**1**

De data aceasta obținem un protocol sistematizat numit **diagrama în linii.** Putem acum observa că toți cei 30 de subiecți *(n=30)* au fost repartizați crescător în funcție de înălțimea lor şi putem afirma că cei mai scunzi subiecți au 171 cm iar cel mai înalt 193 cm. De asemenea, cei mai mulți subiecți au 172 cm şi 177 cm. Iată aşadar că această diagramă de aduce un plus de informație în raport cu şirul nesistematizat prezentat anterior.

Această sistematizare a unui protocol reprezintă o **procedură statistică liminară** deoarece se află la limita analizei statistice pregătind aplicarea tehnicilor ulterioare de prelucrare a datelor.

Diagrama în linii reprezintă o metodă intuitivă de reprezentare grafică a datelor. Următorul pas este acela prin care vom număra liniile trasate în vederea stabilirii **efectivului (frecvența absolută)** fiecărei valori din ansam-

blul rezultatelor. Observăm că efectivul valorii 171 cm este 2 subiecți, adică doi subiecți au înălțimea de 171 cm, efectivul valorii 172 cm este 4 subiecți, patru subiecți au înălțimea 172 cm şi aşa mai departe. Stabilirea efectivelor prin acest procedeu este posibilă în momentul în care amplitudinea de variație este relativ mică (de exemplu pentru notele şcolare, de la 1 la 10). La o amplitudine mai mare acest procedeu este extrem de laborios şi inutil deoarece variabilitatea distribuției devine mare şi se pierde viziunea de ansamblu. Deja, în exemplul nostru, constatăm o serie de tendințe, dar care sunt însă abia conturate. În plus, găsim multe valori pentru care efectivul este zero. Prin urmare, ar fi mai util să facem distribuția mai compactă prin **gruparea valorilor în clase,** astfel încât să putem releva regularitatea ansamblului de date. Vom alege, spre exemplu, gruparea datelor din patru în patru astfel:

171, 172, 173, 174 sau 171 – 174

175, 176, 177, 178 sau 175 – 178

179, 180, 181, 182 sau 179 – 182

183, 184, 185, 186 sau 183 – 186

187, 188, 189, 190 sau 187 – 190

191, 192, 193 sau 191 – 193

Au rezultat un număr de şase clase care cuprind, fiecare, patru valori, ultima clasă fiind formată din trei valori. Distribuția de mai sus va deveni:

Iată că acum lucrurile sunt ceva mai clare. Observăm că cei mai mulți subiecți au înălțimea cuprinsă între 171 cm şi 174 cm în timp ce destul de puțini subiecți pot fi considerați ca fiind înalți. Este mai uşor de sesizat acum imaginea de ansamblu, modalitatea în care se orga-

Tabel 2.2 – Diagrama în linii şi efectivul pentru date grupate în clase

Clasa	Diagrama în linii	Efectiv
171 – 174	||||||||	9
175 – 178	|||||	5
179 – 182	||||||	6
183 – 186	||||	4
187 – 190	|||	3
191 -193	|||	3

nizează datele. Un asemenea procedeu pune în evidență, compact, regularitatea de ansamblu a datelor.

La acest nivel va trebui să facem o serie de precizări deoarece au apărut două concepte noi: conceptul de clasă şi conceptul de interval de grupare. În cazul nostru, numărul de clase a fost stabilit la şase. **Intervalul de grupare (*i*)** se referă la numărul de valori pe care le poate conține o clasă, cu alte cuvinte lărgimea unei clase. Practic, intervalul de grupare se exprimă ca raport dintre amplitudinea de variație şi numărul claselor. Dacă notăm intervalul de grupare cu *„i"* şi numărul de clase cu *„k"*, avem formula:

$$i = \frac{X_{max} - X_{min}}{k} \qquad \text{(formula 2.1)}$$

în cazul nostru formula devenind $i = \frac{193-171}{6} = \frac{22}{6} = 3{,}67 \cong 4$

În exemplul anterior am rotunjit intervalul de grupare *„i"* la valoarea 4. Datorită faptului că înălțimea este o variabilă continuă, ar fi corect să reprezentăm clasele şi astfel:

171,00 – 174,67 *(171,00 + 3,67 = 174,67)*

174,68 – 178,35 *(174,68 + 3,67 = 178,35)*

178,36 – 182,03 *(178,36 + 3,67 = 182,03)*

182,04 – 185,71 *(182,04 + 3,67 = 185,71)*

185,72 – 189,39 *(185,72 + 3,67 = 189,39)*

189,40 – 193,07 *(189,40 + 3,67 = 193,07)*

Totuşi, de ce ne-am rezumat la şase clase şi de ce nu am folosit opt clase sau 3 clase? Care a fost argumentul pe baza căruia am decis utilizarea unui număr de 6 clase? Există vreo metodă prin care putem stabili numărul de clase necesare sau le alegem noi arbitrar?

Atunci când numărul de măsurători *(n)* nu este prea mare, numărul claselor *(k)* poate fi calculat pe baza ***formulei Hahn-Shapiro*** (Vasilescu, 1992)***:***

$$k = INT\left(\frac{n}{5}\right) \quad \text{(formula 2.2)}$$

unde INT înseamnă „partea întreagă”.

În cazul nostru, numărul subiecților *(n)* este 30. Dacă împărțim 30 la 5 obținem valoarea 6. Evident că partea întreagă a lui 6 este tot 6, pe baza acestei formule stabilindu-se numărul optim de clase în care pot fi grupate valorile. Cunoscând numărul claselor am stabilit, după cum s-a precizat mai sus, intervalul de grupare.

Dacă volumul măsurătorilor este mare se utilizează **specificațiile lui Kolker** referitoare la numărul de clase:

Tabelul 2.3 – Specificațiile lui Kolker referitoare la alegerea numărului de clase

n	25-40	41-60	61-99	100	101-160	161-250	251-400	401-630	631-1000	> 1000
k	6	7	8	10	11	12	13	14	15	16-20

Din considerente practice, numărul claselor nu va depăşi valoarea 20.

O alte metodă prin care putem stabili intervalul de distribuție dar şi numărul claselor cunoscând amplitudinea de variație şi numărul de măsurători o reprezintă utilizarea **formulei lui Sturges** (Vasilescu, 1992)**:**

$$\boldsymbol{i = \frac{X_{max} - X_{min}}{1 + 3{,}322 \times lg(n)}} \quad \text{(formula 2.3)}$$

În situația noastră, logaritmul zecimal din 30 este 1,47. În mod curent putem găsi logaritmul zecimal în tabelele de logaritmi sau utilizând un calculator ştiințific. Cunoscând această valoare, intervalul de distribuție devine:

$$i = \frac{193 - 171}{1 + 3{,}322 \times 1{,}47} = \frac{22}{5{,}88} = 3{,}74 \cong 4$$

Observăm că valoarea obținută aici se apropie foarte mult de valoarea stabilită prin formula anterioară. Această formulă poate fi folosită pentru valori destul de mari ale lui *n* respectând însă limita de 20 de clase.

Referitor la numărul claselor şi la intervalele de grupare se impun următoarele observații:

1. Dacă alegem intervale de grupare prea mici, gruparea datelor scoate în evidență lucruri de amănunt ce estompează tendința de ansamblu a datelor şi deci a procesului pe care îl studiem;
2. Intervalele de grupare prea mari pot estompa iregularitățile distribuției şi atenuează specificul fenomenului studiat, pierzându-se precizia cu care identificăm o valoare individuală;
3. În mod obişnuit, vom utiliza, dacă măsurătoarea s-a făcut folosind numere întregi, limite de interval exprimate tot în numere întregi. Limita inferioară a intervalului superior va trebui să fie mai mare cu o unitate decât limita superioară a intervalului inferior. De exemplu:

171 – 174 *şi nu* *171 – 174*

175 – 178 *174 – 178*

deoarece nu vom şti unde să încadrăm valoarea 174

La fel, *171,00 – 174,67* *şi nu* *171,00 – 174,67*

174,68 – 178,35 *174,67 – 178,35*

După cum putem însă observa, măsurătorile inițiale s-au efectuat în numere întregi. Un exces de precizie la acest nivel este exagerat şi poate produce confuzii. Un subiect cu înălțimea de 174 unde va fi inclus? În primul

caz, evident, în prima clasă. În al doilea caz însă, poate fi inclus şi în prima şi în a doua clasă. **De aceea niciodată precizia claselor nu va depăşi precizia măsurătorilor efectuate.**

Gruparea în clase fiind încheiată, obţinem aşa-numita **distribuţie statistică** ce formează baza prelucrărilor de date ulterioare. În exemplul nostru vom avea următoarea distribuţie statistică:

Tabelul 2.4 – Distribuţia statistică pe clase

Clasa	Distribuţia în linii	f_a	$f_{c\uparrow}$	$f_{c\downarrow}$	%	‰
171 – 174	||||||||	9	9	30	30,00	300,00
175 – 178	|||||	5	14	21	16,66	166,66
179 – 182	||||||	6	20	16	20,00	200,00
183 – 186	||||	4	24	10	13,33	133,33
187 – 190	|||	3	27	6	10,00	100,00
191 -193	|||	3	30	3	10,00	100,00
		30				

Însumând efectivele (frecvenţele) tuturor intervalelor, va rezulta, evident, numărul total de măsurători *(n)*. Este de asemenea posibil să calculăm **frecvenţele cumulate,** fie ascendent, fie descendent, ca expresie a adunări frecvenţei intervalului curent la frecvenţa cumulată a intervalului anterior. Astfel, pentru primul interval, atât frecvenţa cât şi frecvenţa cumulată ascendentă au valoarea 9. Pentru al doilea interval frecvenţa absolută este 5 iar frecvenţa cumulată ascendentă devine 9+5=14. Al treilea interval are frecvenţa absolută 6 iar frecvenţa cumulată ascendentă 14+6=20 şi aşa mai departe. Procedura de calcul este analoagă şi în cazul frecvenţelor cumulate descendent, cu singura deosebire că adunarea se face începând de la ultimul interval.

Rolul frecvenţelor cumulate este acela de a „ordona" crescător sau descrescător datele grupate în intervale şi de a indica numărul de cazuri, câţi subiecţi se află până la o anumită valoare (frecvenţele cumulate ascendente) sau câţi subiecţi au depăşit o anumită valoare (frecvenţele cumulate descendente). Dacă discutăm de ordonare, este evident că variabila trebuie să se afle cel puţin la un nivel ordinal de măsură, calculul frecvenţelor cumulate în cazul variabilelor nominale fiind un nonsens.

Putem exprima, de asemenea, **frecvențele relative**, ca proporție a scorurilor dintr-o clasă. Cele mai utilizate frecvențe relative sunt procentele (%) şi promilele (‰). Ca modalitate de calcul al procentelor sau promilelor vom folosi regula de trei simplă, astfel:

Pentru prima clasă vom avea următoarea valoare:

dacă 30 reprezintă 100%

atunci 9 reprezintă x

$$de\ unde\ x = \frac{9 \times 100}{30} = \frac{900}{30} = 30\%$$

II.2 Reprezentarea grafică a datelor

Sintagma „o imagine face mai mult decât o mie de cuvinte” este valabilă şi în statistică. Uneori este mai convenabil să reprezentăm grafic datele pentru a ne forma intuitiv o reprezentare a acestora. Nu trebuie să uităm însă faptul că în statistică afirmațiile pe care le facem trebuie demonstrate iar demonstrația se poate face doar cifric.

Există multe moduri prin care putem reprezenta grafic datele pe care le vom discuta în momentul în care tematica le va impune. Pentru moment ne vom rezuma la graficele cu bare, histograma, poligonul frecvențelor, poligonul frecvențelor cumulate şi plăcinta.

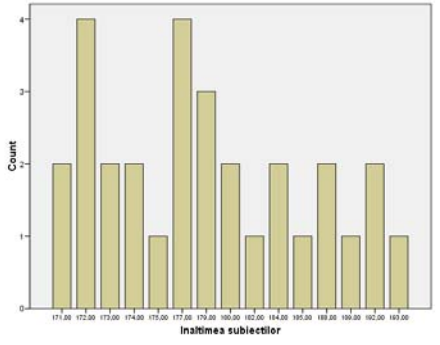


Figura 2.1– Graficul cu bare al înălțimii subiecților

Reprezentarea prin graficele cu bare constă în ridicarea unei perpendiculare de lungime egală cu frecvența pentru fiecare valoare sau clasă aşezată pe axa O_x. Cu ajutorul barelor putem reprezenta atât datele negrupate cât şi datele grupate. Astfel, pentru exemplul nostru, prezentăm reprezenta-

rea prin grafice cu bare a înălțimii subiecților sub formă de valori. Un tip particular de reprezentare a graficelor cu bare este **histograma**. O histogramă se deosebeşte de un grafic cu bare prin aceea că, în locul liniilor, sunt ridicate dreptunghiuri cu baza egală cu intervalul folosit. Pentru a putea folosi histogramele, datele trebuie să fie grupate aşadar pe intervale egale.

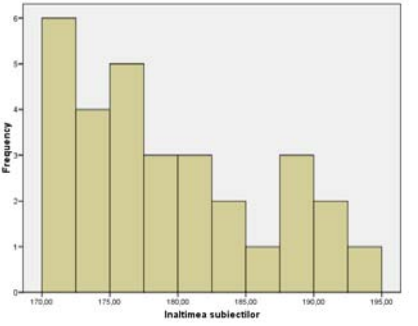


Figura 2.2 – Histograma pentru înălțimea subiecților

Dacă unim prin linii vârfurile graficelor cu bare sau mijlocul bazei de sus a histogramelor obținem un alt tip de grafic şi anume **poligonul frecvențelor absolute sau relative.** Poligonul frecvențelor absolute se referă la exprimarea cifrică directă a efectivelor în timp ce poligonul frecvențelor relative conține exprimarea procentuală a acestora.

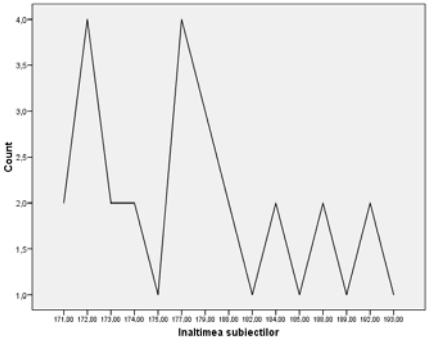


Figura 2.3 – Poligonul frecvențelor absolute

De asemenea este posibilă construirea **poligonului frecvențelor cumulate** ca reprezentare grafică a funcției de repartiție cumulativă. La fel ca şi în cazul graficului cu bare, în dreptul fiecărei valori se înalță o perpendiculară de mărime egală cu frecvența cumulată. Unind vârfurile acestor perpendiculare se obține poligonul frecvențelor cumulate care poate fi exprimat absolut sau relativ.

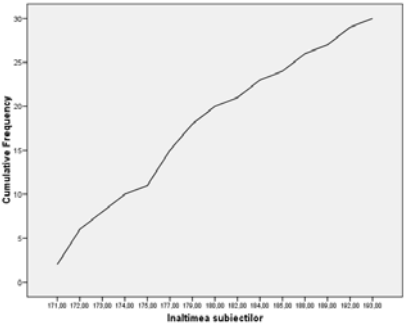


Figura 2.4 – Poligonul frecvențelor cumulate

Uneori este mai convenabil să reprezentăm, mai ales atunci când lucrăm cu procente, compoziția măsurătorilor printr-o structură radială. Un asemenea grafic se numește **plăcintă (pie)** și este util atunci când ne interesează să urmărim proporția cu care fiecare clasă participă la construcția „întregului".

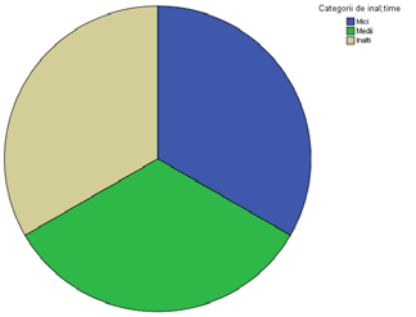


Figura 2.5 – Reprezentarea prin diagramă de structură (plăcintă)

Pentru o variabilă continuă, deoarece avem o densitate de probabilitate pe intervalul de valori, reprezentarea grafică se face sub forma unei curbe numită **curbă de distribuție**, însă acest aspect va face obiectul unui capitol ulterior.

În concluzie:

- *O colecție de date are o valoare informativă extrem de redusă. În vederea extragerii unor informații preliminare este necesară ordonarea și gruparea datelor;*
- *La o amplitudine de variație mică pot fi ordonate valorile și numărate efectivele fiecărei valori realizându-se astfel un protocol sistematizat;*
- *La o amplitudine de variație mare și la măsurători numeroase se va realiza gruparea datelor în clase, pe baza unui interval de grupare. Atât numărul claselor cât și intervalul de grupare pot fi calculate în baza unor formule sau a unor tabele de referință;*
- *O distribuție statistică reprezintă un protocol de date sistematizat și organizat care conține clasele, diagramele în linii (puncte, steluțe etc.) și frecvențele absolute. Alte informații care mai pot fi incluse sunt frecvențele relative (procentuale), frecvențele cumulate absolute și relative;*
- *Distribuțiile statistice pot fi reprezentate și grafic folosindu-se graficele cu bare, histogramele, poligoanele de frecvențe sau graficele plăcintă.*

Întrebări și exerciții:

- *La un test de inteligență subiecții au obținut următoarele scoruri: 16, 20, 20, 9, 17, 6, 11, 5, 15, 10, 5, 16, 16, 11, 11, 13, 14, 13, 10, 13, 13, 13, 15,*

12, 21, 7, 14, 24, 13, 19, 15, 17, 14, 8, 8, 17, 15, 9, 4, 22, 12, 23, 10, 18, 8, 13, 12, 14, 11, 13, 14. Realizaţi distribuţia statistică precizând toţi indicatorii studiaţi şi reprezentaţi grafic.

- *Ce diferenţe sunt între frecvenţa absolută şi frecvenţa relativă?*
- *Ce rol au frecvenţele cumulate?*
- *Când poate fi folosit un grafic de tip „plăcintă”?*
- *În câte clase reprezentaţi un volum de n=32.458 de cazuri?*
- *Ce este o distribuţie statistică?*
- *Precizaţi avantajele şi dezavantajele distribuţiei statistice organizată pe clase.*

III. STATISTICI DESCRIPTIVE – TENDINȚA CENTRALĂ

În acest capitol se va discuta despre:

- Conceptul de statistici descriptive;
- Definirea noțiunilor de populație și eșantion;
- Indicatori ai tendinței centrale – media, mediana și modul;
- Obținerea indicatorilor tendinței centrale în SPSS;

După parcurgerea capitolului, cursanții vor fi capabili să:

- Înțeleagă semnificația statisticilor descriptive;
- Facă distincția între populație și eșantion;
- Distingă între diferitele modalități de eșantionare;
- Calculeze indicatorii tendinței centrale;
- Utilizeze corect indicatorii tendinței centrale în funcție de modul de distribuție al datelor;
- Să extragă indicatorii tendinței centrale utilizând programul SPSS.

Statistica descriptivă se referă la totalitatea metodelor care permit descrierea şi gruparea în diferite moduri a datelor rezultate din cercetări. Orice procedeu de analiză a datelor presupune efectuarea unor operațiuni preliminare de statistică descriptivă. Statisticile descriptive nu au ca scop efectuarea unor predicții ci doar sumarizarea şi prezentarea datelor. Procedeele utilizate în statisticile descriptive se referă la gruparea şi prezentarea datelor în tabele şi grafice, calculul indicatorilor tendinței centrale şi ai variabilității, indicatori ai asocierii şi legăturii dintre variabile (studii corelaționale simple).

Încă de la început s-a stabilit că, prin statistică, putem descrie, compara şi relaționa variabile. Totodată, modul de realizare al unei variabile se face prin atribuirea de valori rezultate din diferite surse de date. Totuşi, cum culegem aceste date? Cum facem ca pornind de la un număr relativ redus de su-

biecți să extindem cunoştinţele noastre la nivelul întregii populații? Şi în definitiv ce înseamnă populație şi ce înseamnă eşantion?

III.1 Populație şi eşantion

Când discutăm de populație, nu este obligatoriu să ne referim la populația unei țări sau a unui oraş; în general, ne referim la un grup distinct de persoane sau obiecte. De exemplu, vorbim de populația fumătorilor, populația deficienților mintali, populația consumatorilor de Coca-Cola sau populația maşinilor Lamborghini. Acest lucru se referă la **totalitatea obiectelor, de un anumit tip, existente într-un spațiu sau teritoriu, la un moment dat.** O primă observație se referă la faptul că populația este relativă la un criteriu (de exemplu cel teritorial) sau la mai multe criterii (de exemplu automobile şi România). Prin urmare, când vorbim de populație relevantă, avem în vedere totalitatea obiectelor care satisfac unul sau mai multe criterii.

De exemplu, dacă desfăşurăm un studiu în care dorim să demonstrăm eficiența unui medicament ce va accelera abandonul fumatului în România, ne interesează criteriul teritorial (România) dar şi calitatea de fumător, prezentă sau trecută.

Populația este definită deci, ca fiind totalitatea „unităților de informație" care constituie obiectivul de interes al unei investigații. Întotdeauna, la începutul unei cercetări, trebuie menționată populația cercetării în vederea specificării domeniului la care vom extinde rezultatele, concluziile obținute.

De cele mai multe ori nu putem studia întreaga populație din cauza numeroaselor limitări pe care le-ar impune un asemenea demers (costuri, timp, accesibilitate, dinamica populației etc.). De aceea majoritatea cercetărilor se efectuează pe un grup de subiecți extraşi din populație, care să prezinte caracteristicile acesteia, grup mult redus însă din punct de vedere numeric. Un asemenea grup poartă numele de **eşantion**. Un eşantion este deci o selec-

ție a indivizilor dintr-o populație. Cercetătorii utilizează eşantioane, deoarece lucrul cu acestea este mai ieftin, mai rapid şi mai uşor în comparație cu cercetarea întregii populații iar rezultatele pot fi extinse la nivelul întregii populații.

Eşantionul reprezintă „unitățile de informație" selecționate pentru a fi efectiv studiate. Ideea pe care se bazează cercetările orientate pe eşantioane, este aceea că se pot face aprecieri asupra unei întregi populații, în anumite condiții, doar pe baza caracteristicilor măsurate pe o parte a acesteia.

Totuşi, pentru ca rezultatele obținute prin studiul eşantionului să poată fi extinse la nivelul întregii populații, acel eşantion trebuie să fie reprezentativ, adică să întrunească caracteristicile de bază ale populației din care a fost extras.

Să presupunem, de exemplu, că fierbem o oală de fasole. Cum verificăm dacă fasolele au fost fierte? În primul rând, amestecăm fasolele în oală. Dacă nu am amesteca fasolele, am putea risca să le luăm pe cele de la suprafață (care sunt mai puțin fierte) nu şi pe cele de la fund (care probabil sunt mai fierte). După aceea, luăm o lingură de fasole şi gustăm. În funcție de rezultat, decidem dacă acestea sunt sau nu sunt fierte. Nu trebuie se mâncăm toată oala de fasole ca să decidem dacă sunt sau nu fierte. Este suficient să verificăm un eşantion de fasole pentru ca mai apoi să extindem cunoştința la nivelul întregii oale. Iată deci că exact acesta este şi principiul eşantionării. Prin amestecarea în oala de fasole am creat posibilitatea ca fiecare boabă de fasole să aibă şansa de a fi aleasă. Apoi, am extras aleatoriu un eşantion de boabe de fasole pe care l-am verificat. Dacă am decis că sunt fierte, nu voi mai continua cercetarea. Şansele să mă înşel sunt prea mici[5].

Afirmam anterior că un eşantion trebuie să fie reprezentativ, adică să îndeplinească caracteristicile populației din care face parte. Evident, în mo-

[5] Exemplul a fost preluat şi adaptat după Lungu Ovidiu – Ghid introductiv pentru SPSS 10.0 pentru Windows

mentul în care voi dori să ştiu dacă sunt fierte fasolele, voi extrage un eşantion de boabe de fasole şi nu unul de boabe de mazăre din oala alăturată. Posibilitatea de a generaliza cunoştinţele dobândite la nivelul eşantionului către întreaga populaţie este vitală în cercetarea experimentală. De aceea, asigurarea reprezentativităţii unui eşantion este de maximă importanţă.

Metoda prin care selectăm un eşantion dintr-o populaţie are implicaţii asupra validităţii şi generalizării concluziilor dobândite în urma studiului eşantionului. Metodele de eşantionare se împart în două mari categorii: metode prin care eşantionul este extras aleatoriu dintr-o populaţie şi metode prin care eşantionul nu este extras aleatoriu.

Într-un eşantion care nu este extras aleatoriu, probabilitatea ca un subiect să facă parte din eşantion nu poate fi calculată. Este exemplul clasic în care subiecţii sunt selectaţi pe bază de voluntariat sau cei care returnează chestionarele completate prin poştă.

Într-un eşantion extras aleatoriu, fiecare individ are o probabilitate calculabilă de a fi inclus în eşantion iar această probabilitate va trebui să fie egală pentru fiecare subiect. Este cazul extragerii unui eşantion dintr-o bază de date care conţine, aleatoriu, întreaga populaţie.

Orice eşantion se stabileşte în baza unei **scheme de eşantionare**. O schemă de eşantionare reprezintă un set de tehnici şi reguli în baza cărora din populaţie se extrage eşantionul necesar iar compoziţia acestuia este aleatoriu derivată în funcţie de definiţiile probabilistice ale schemei de eşantionare.

III.1.1 Eşantionarea simplu randomizată

Randomizare nu înseamnă hazard absolut ci este o metodă prin care fiecare individ are o şansă calculabilă de a fi ales în eşantion. Prin randomizare simplă, fiecare individ are şanse egale să fie selectat. Pentru a efectua o asemenea eşantionare, avem nevoie de o **bază de eşantionare**, o listă a întregii populaţii (de exemplu, lista exhaustivă a fumătorilor din România). Fieca-

re individ din acea listă are asociat un număr. Se folosesc apoi tabelele de numere aleatoare sau numerele aleatoare se generează de un computer. Practic, numărul aleatoriu generat reprezintă numărul de ordine al individului de pe listă. Procedura continuă până când s-a extras întregul eşantion. Metoda randomizării este o metodă foarte simplă însă dificultatea constă în posibilitatea obținerii unei liste exhaustive pe criteriile dorite.

III.1.2 Randomizarea pe cote

Este o metodă pseudo-randomizată de selectare a unui eşantion în care se iau în considerare anumite caracteristici ale populației cum ar fi sexul, vârsta, profesia etc. În eşantion se intenționează respectarea proporției în care aceste caracteristici se regăsesc în populație. După ce se identifică proporția fiecărei caracteristici ce va fi inclusă în eşantion, fiecărui evaluator îi revine sarcina să evalueze un anumit număr de persoane ce prezintă acea caracteristică, alegerea persoanelor rămânând la latitudinea acestuia. Este important ca acea caracteristică să poată fi identificată cu uşurință şi fără ambiguitate de către evaluator, altminteri putând apărea serioase probleme la nivelul reprezentativității. De exemplu, dacă avem următoarea structură a populației din România după două criterii – vârstă şi sex:

Tabel 3.1 – Structura populației

Vârsta (ani)	Bărbați	Femei
15-20	10000	9000
21-30	12000	13000
31-50	12000	14000
peste 50	12000	18000

O eşantionare pe cote presupune respectarea proporției în care se găseşte populația generală, la aceste două criterii şi în eşantion. Astfel, dacă extragem un eşantion de 1000 de persoane, proporția subiecților din acest eşantion va respecta proporția subiecților din populația generală. Prin urmare, vom obține:

Tabel 3.2 – Proporții eşantion

Vârsta (ani)	Bărbați	Femei
15-20	10%	9%
21-30	12%	13%
31-50	12%	14%
peste 50	12%	18%

Acesta este principiul eşantionării pe cote. Un evaluator va administra instrumentele de evaluare unui număr de, să spunem, 120 de bărbați cu vârsta

cuprinsă între 31 şi 50 de ani, având libertatea de a alege singur persoanele, atâta timp cât respectă aceste două criterii.

O problemă apare în momentul în care unul dintre criterii are o ponderе extrem de mică în populație (de exemplu, când studiul se referă la persoane cu un IQ sub 70). În acest caz, numărul persoanelor selectate în eşantion s-ar putea să fie prea mic pentru a permite generalizarea. Într-o asemenea situație se procedează la „supra-eşantionare" adică la mărirea acestui număr. Se consideră că pentru anumite caracteristici foarte rare, acest fapt nu afectează reprezentativitatea eşantionului.

III.1.3 Eşantionarea stratificată

În condițiile în care putem identifica o serie de straturi după care este împărțită populația în funcție de unul sau mai multe criterii, vom putea proceda la o eşantionare randomizată pentru fiecare strat. Acest lucru ne permite să ne asigurăm că fiecare grup este reprezentat în eşantionul final. Numărul fiecărui grup este proporțional cu mărimea stratului, însă dacă acel grup este foarte mic, se procedează, la fel ca în cazul anterior, la mărirea numărului de subiecți în eşantion pentru asigurarea reprezentativității. De exemplu, putem depista stratul fumătorilor, stratul nefumătorilor şi stratul celor care au renunțat la fumat. Aflăm care este ponderea fiecărui strat în populația generală şi apoi, din fiecare strat, extragem aleatoriu eşantionul nostru. Metoda se aseamănă cu cea prezentată anterior, diferența fiind aceea că în eşantionarea pe cote evaluatorul poate alege liber subiecții dintr-un grup, în timp ce aici subiecții din grup sunt aleşi dinainte în baza eşantionării randomizate simple. Dacă discutăm de eşantionare randomizată simplă, evident că se ridică din nou problema bazei de eşantionare. Deşi este mai precisă în comparație cu eşantionarea pe cote, eşantionarea stratificată se loveşte de aceeaşi problemă, dificultatea obținerii bazei de eşantionare.

III.1.4 Eşantionarea pe cluster

Un cluster reprezintă o unitate care prin ea însăşi se prezintă ca un individ dintr-o populație. Spre exemplu, dacă populația o considerăm ca fiind cea a liceelor din România, un cluster este liceul „C Negruzzi" din Iaşi, un alt cluster liceul „Mircea cel Bătrân" din Constanța şi aşa mai departe. În acest caz, eşantionarea se face pe unități de eşantionare iar evaluarea se face exhaustiv la nivelul clusterului. Aşadar, eşantionul este stabilit pe unități de eşantionare şi nu pe indivizi iar la nivelul unei unități de eşantionare se evaluează exhaustiv toate persoanele. Spre exemplu, dacă baza de eşantionare o constituie liceele din România care sunt, să presupunem, în număr de 3000, atunci vom extrage aleatoriu un eşantion de 30 de licee pe care le vom investiga exhaustiv. Un asemenea liceu joacă rolul unui individ dintr-o eşantionare randomizată şi poartă numele de **unitate de eşantionare**.

Există şi alte modalități de eşantionare, însă acestea sunt cele mai importante pentru cercetarea în ştiințele socio-umane. Rămâne însă problema cât de mare să fie eşantionul pentru a asigura generalizarea concluziilor la nivelul întregii populații. Pentru a trata acest aspect, va trebui să ştim câteva lucruri despre distribuția datelor într-o populație, mai precis despre curba de distribuție normală. De aceea, asupra acestei probleme vom reveni şi o vom trata în amănunt într-un alt capitol.

III.2 Indicatori ai tendinței centrale

Calitatea datelor colectate este fundamentală pentru scopul cercetării. Am aflat în capitolul al doilea cum putem organiza setul de date brute astfel încât informațiile să devină inteligibile. Urmează să ne asigurăm că metoda prin care am colectat datele este adecvată şi că datele ne pot informa într-o manieră ştiințifică. Am spus, de asemenea, că pentru volume mari de date, acest proces de organizare poate fi unul extrem de laborios şi mare consumator de timp. De asemenea, probabil că ați fost surprinşi că în capitolul al doilea nu avem nici o aplicație în SPSS. Motivul pentru care nu am inclus nici o

aplicație practică a fost acela că, în realitate, în momentul în care lansăm o procedură de analiză de date, SPSS face automat această organizare. Indiferent cât de sofisticate ar fi analizele de date pe care le efectuăm, nu putem obține la „ieşire" decât ceea ce am introdus la „intrare". Iată deci că înainte de a intra în procedee statistice mai elaborate am discutat câteva elemente referitoare la eşantionare. Ceea ce trebuie să rețineți, este faptul că o cercetare greşit proiectată duce în mod sigur la rezultate eronate.

Organizarea şi sistematizarea datelor şi reprezentarea grafică a acestora nu sunt suficiente în cele mai multe cazuri. Avem nevoie de mai mult decât de desene şi grafice frumoase pentru a da un sens acelor date şi pentru a extrage concluzii utile. Uneori avem nevoie să stabilim relațiile dintre două sau mai multe variabile sau să vedem cum scorul la o anumită variabilă este influențat de scorurile altor variabile. Alteori trebuie să comparăm variabilele, să vedem dacă sunt diferențe între ele şi care este sensul acestor diferențe. Toate aceste aspecte sunt reunite sub denumirea generică de **analiză a datelor.**

În al doilea capitol am văzut deja cum datele se concentrează în jurul unor valori şi am constatat că în raport cu dimensiunea eşantionului, subiecții pot fi considerați mai degrabă scunzi decât înalți. Iată că prin acest procedeu putem intui că datele noastre tind spre anumite valori considerate ca fiind centrale. Aşadar, putem vorbi despre o **tendință centrală** a rezultatelor.

Prima şi cea mai comună formă de descriere a datelor o reprezintă analiza tendinței centrale. Măsura tendinței centrale ne dă o indicație asupra scorurilor tipice din colecția de date. Cu alte cuvinte, ne indică modul în care se grupează datele în jurul unei valori. Cei mai importanți indicatori ai tendinței centrale sunt media, mediana şi modul.

Procedeele statistice urmăresc, în aproape toate cazurile, modelarea lumii reale. Un model statistic reprezintă o reprezentare probabilistică a unei realități care nu poate fi cunoscută nemijlocit. Ar fi foarte greu să cunoaştem

structura geologică a unui munte. Nu putem aduce muntele în laborator să-l studiem şi nici nu putem trimite mii de geologi care să investigheze fiecare rocă din compoziția muntelui. Totuşi, ştim că muntele este calcaros sau granitic, putem aprecia vechimea acestuia şi era geologică de formare. Acest lucru devine posibil prin studiul unor eşantioane prelevate din munte pe baza cărora s-a construit un model probabilistic al muntelui.

III.2.1 Media

Considerăm că termenul de medie – mai precis media aritmetică – este conceptul cel mai uşor de înțeles din întreaga statistică. Au fost nenumărate situațiile în care ați întâlnit media: media la matematică pe trimestru sau semestru să vedeți dacă ați rămas corigent sau ați luat premiu, media de la bacalaureat care vă ridică sau vă coboară şansele de admitere la facultate, la fel ca şi media anilor de studiu, media de vârstă din familia dumneavoastră sau media cheltuielilor zilnice care vă ajută să vă planificați mai bine bugetul.

Deci ce este media? Media nu este altceva decât suma valorilor unei variabile raportată la numărul măsurătorilor, fiind poate cel mai simplu model statistic. Aceasta este **media aritmetică**, deoarece în statistică mai discutăm şi de media geometrică, media caracteristicilor alternative, media pătratică, media rangurilor etc. Aceste concepte le întâlnim însă mai rar în domeniul ştiințelor socio-umane şi prin urmare nu vom face decât să le amintim.

Probabil că vi se va părea ciudată afirmația anterioară. Cum adică media este un „model statistic”? Ei bine, aşa este, deoarece nu reprezintă altceva decât o valoare ipotetică ce poate fi obținută din orice set de date măsurate la un nivel parametric. Dacă, de exemplu, înregistrăm numărul de frați pe care îi au cinci prieteni de-ai mei, am obține următoarele valori: primul prieten are un frate, al doilea 2 frați, al treilea şi al patrulea 3 frați şi al cincilea 4 frați. Câți frați au în medie cei cinci prieteni ai mei? Dacă adunați cele cinci valori şi împărțiți la cinci, obținem o medie de 2,6 frați. Câți frați??? Repet câți frați? Media spune clar: 2,6 frați. Ce înseamnă 2,6 frați? Probabil doi

fraţi întregi şi unul căruia îi lipseşte o mână sau un picior. Iată de ce media nu este altceva decât un model ipotetic, un model statistic care reprezintă tendința, direcția spre care converg datele.

Când vorbim de medie, trebuie să înțelegem şi condițiile în care o putem folosi, deoarece utilizarea ei în condiții inadecvate este inutilă şi primejdioasă, putând duce la multe confuzii şi erori de interpretare. Am spus deja că putem folosi media doar în cazul în care datele se află cel puțin la un nivel de măsurare de interval, altminteri nu are sens să discutăm despre medie. Nu putem face, vă amintiți, media între un colonel şi un căpitan. Ce-ar rezulta??!! În nici un caz un maior. De asemenea, media poate fi folosită doar în cazurile în care valorile individuale se grupează în jurul acesteia, iar valorile care se abat de la medie se anulează reciproc. Cu alte cuvinte, suma valorilor individuale mai mari decât media şi suma valorilor mai mici decât media tind să se anuleze. Vom vedea, în acest capitol, care sunt limitele mediei şi cum le putem evita şi atunci veți înțelege mai bine aceste concepte.

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i}{n} \qquad \text{(formula 3.1)}$$

Iată şi prima formulă care are darul să vă „sperie”. În realitate această formulă este formula mediei aritmetice menționată mai sus. S-ar traduce prin suma tuturor valorilor de la *i* la *k* împărțită la numărul valorilor *n*. Semnul sigma este un semn folosit pentru a exprima suma. Observăm că pentru a avea sens, trebuie să existe cel puțin un element. Până acum nimic avansat. Media aritmetică se mai notează cu *m* sau folosind $\overline{x}$ ca în formula noastră. De fapt, dacă respectăm strict convențiile de notare din literatura de specialitate, atunci când vorbim de o populație, vom nota media cu μ *(miu)* iar când de referim la un eşantion o vom nota cu *m* sau $\overline{x}$.

Iată, avem un număr de 10 măsurători ale unor scoruri brute la un chestionar de anxietate, după cum urmează: 10, 22, 31, 9, 24, 27, 29, 9, 23,

12. Media aritmetică va fi suma acestora împărțită la numărul lor, în cazul nostru 10. Iată că efectuând calculele obținem media de 19,6. Este greu, nu?

$$\overline{x} = \frac{10 + 22 + 31 + 9 + 24 + 27 + 29 + 9 + 23 + 12}{10} = \frac{196}{10} = 19{,}6$$

Vom reconsidera acum exemplul din capitolul al doilea şi vom încerca să-l utilizăm pentru a efectua calculele în vederea stabilirii celor trei indicatori ai tendinței centrale. Vi-l reamintesc:

Exemplu: *Presupunem că măsurăm înălțimea unor subiecți în centimetri şi obținem următoarele rezultate: 182, 175, 174, 189, 177, 177, 180, 173, 188, 171, 184, 192, 180, 172, 177, 193, 184, 192, 172, 177, 174, 171, 172, 179, 185, 179, 183, 172, 179, 188.*

Pentru a calcula media, va trebui să adunăm toate aceste 30 de valori şi să împărțim suma rezultată la 30 care reprezintă numărul măsurătorilor. Începeți să adunați….. Suma tuturor acestor valori este 5391. Dacă împărțim această sumă la 30 vom obține valoarea 179,7 care nu reprezintă altceva decât media înălțimii celor 30 de subiecți.

Ați observat probabil că v-a luat destul de mult timp să calculați această medie folosind adunarea valorilor. Nu ați avut decât 30 de valori de adunat. Acum imaginați-vă că aveți 3000 sau 5000 de valori. Vă puteți da cu uşurință seama că în asemenea condiții calculul mediei devine extrem de dificil şi foarte susceptibil la erorile ce pot să apară. Când volumul datelor este destul de mic, putem calcula media prin însumarea valorilor şi raportarea acestei sume la numărul de măsurători. Dacă volumul de date este însă mare iar datele sunt grupate, vom prefera o altă metodă de calcul folosind aceste date grupate. Iată o primă aplicație a datelor grupate. Vă mai aduceți aminte de distribuția statistică folosită în capitolul doi? O vom reproduce şi apoi vom modifica acest tabel astfel încât să ne permită calculul mediei. Ştim că intervalul de grupare este *i=4* iar acest interval a fost stabilit prin aproximare.

Tabelul 3.3 – Distribuţia statistică

Clasa	Distribuţia în linii	f	f_c	%	$\%_c$
171 – 174	||||||||	9	9	30,00	30,00
175 – 178	|||||	5	14	16,66	46,66
179 – 182	||||||	6	20	20,00	66,66
183 – 186	||||	4	24	13,33	80,00
187 – 190	|||	3	27	10,00	90,00
191 -193	|||	3	30	10,00	100,0
		n=30		**100%**	

Urmează să stabilim valoarea centrală x_k ce nu reprezintă altceva decât mijlocul intervalului curent. Astfel, pentru primul interval avem valorile 171, 172, 173, 174. Valoarea centrală a intervalului va fi aşadar între 172 şi 173, mai precis 172,5. Pentru al doilea interval avem 175, 176, 177, 178 şi deci valoarea centrală devine în acest caz 176,5 şi aşa mai departe. Ultimul interval conţine însă doar trei elemente şi anume 191, 192, 193. Aici valoarea centrală va fi deci 192.

Urmează apoi, stabilirea valorii fxx_k care reprezintă înmulţirea, pentru fiecare clasă, a valorii x_k cu frecvenţa absolută f. Pentru primul interval vom avea, aşadar, 172,5 x 9 = 1552,5. Al doilea interval va avea ca rezultat produsul 176,5 x 5 = 882,5 şi aşa mai departe până la ultimul interval.

Adunăm apoi toate aceste produse obţinute în coloana fxx_k pentru a stabili suma produselor, **Σ**. În cazul nostru, 1552,5 + 882,5 + 1083 + 738 + 565,5 + 576 = 5397,5. Constatăm că valoarea este foarte apropiată de valoarea exactă obţinută prin însumarea tuturor valorilor (5391). Acum, dacă împărţim totalul 5397,5 la numărul de măsurători n vom obţine valoarea mediei 179,91. Această nouă valoare diferă foarte puţin de media reală obţinută prin însumarea valorilor. Acest lucru se întâmplă deoarece media pune în evidenţă tendinţa centrală a rezultatelor constatate într-o

Tabel 3.4 – Calculul mediei aritmetice

Clasa	x_k	F	fxx_k
171 – 174	172,5	9	1552,5
175 – 178	176,5	5	882,5
179 – 182	180,5	6	1083
183 – 186	184,5	4	738
187 – 190	188,5	3	565,5
191 -193	192	3	576
		n=30	**Σ = 5397,5**

cercetare iar gruparea datelor introduce o eroare în calculul mediei pentru că se face ipoteza că elementele fiecărei clase sunt concentrate la mijlocul intervalului, ceea ce nu este aşa. Cei zece subiecți din primul interval, ştim bine, nu au înălțimi concentrate în jurul valorii de 172,5 şi cu atât mai puțin subiecții din celelalte intervale. Cu cât intervalul de grupare este mai mare cu atât această eroare de grupare va avea o valoare mai mare. Statisticienii consideră însă că la volume mari de date eroarea este neglijabilă comparativ cu timpul alocat însumării valorilor.

Iată că, din cele expuse mai sus, putem deduce o nouă formulă pentru medie, valabilă în cazul datelor grupate:

$$\overline{x} = \frac{\sum_{i=1}^{k} f_k x_k}{n} \quad \text{(formula 3.2)}$$

unde f_k reprezintă frecvența absolută a unei clase k, x_k reprezintă mijlocul clasei k iar n reprezintă numărul total de subiecți

Calculând media se obține o măsură a nivelului mediu relativ la un eşantion studiat, fapt ce permite apoi comparații între grupe.

Media aritmetică prezintă o serie de **proprietăți importante**, foarte utile înțelegerii condițiilor de utilizare ale acestui indicator statistic.

- Adăugând sau scăzând o constantă la fiecare valoare (scor) individual, media se măreşte sau se micşorează cu valoarea acelei constante;
- Înmulțind sau împărțind fiecare scor individual cu o constantă, media se multiplică sau se divide cu valoarea acelei constante;
- Suma scorurilor cu valori mai mici decât media este egală cu suma scorurilor mai mari decât media, astfel încât scorurile mai mici de medie anulează scorurile mai mari decât media;

O variantă a mediei, utilizată destul de des în cercetările socio-umane, se referă la **media ponderată**. În cazul mediei aritmetice simple, fiecare scor particular are aceeaşi pondere în calculul acestui indicator, situație care se schimbă atunci când ne referim la media ponderată. În acest caz, ponderea scorurilor este diferită. Să presupunem că administrăm unui subiect un număr de 4 teste de inteligență, notate *„testul 1", „testul 2", „testul 3"* şi *„testul 4"* la care subiectul obține următoarele scoruri (pe o scală de la 1 la 10): 6, 8, 9, 7. Media aritmetică se calculează foarte simplu, adunând aceste valori şi apoi împărțind suma rezultată la patru. Obținem un scor mediu de 7,5. Pentru calculul mediei ponderate trebuie ca fiecare dintre cele patru teste să aibă o pondere în scorul final. Primul test, fiind mai greu, va avea o pondere mai mare, să spunem 5. Al doilea test va avea ponderea 1, al treilea test 2 şi al patrulea test 3. Totalul ponderilor va fi, aşadar, 5+1+2+3=11. Cunoscând aceste elemente, putem calcula media ponderată, astfel:

$$\overline{x}' = \frac{(5 \times 6) + (1 \times 8) + (2 \times 9) + (3 \times 7)}{11} = \frac{30 + 8 + 18 + 21}{11} = \frac{77}{11} = 7$$

Având în vedere algoritmul de mai sus, putem scrie în continuare formula mediei ponderate, după cum urmează:

$$\overline{x}' = \frac{\sum_{i=1}^{k} p_i x_i}{\sum_{i=1}^{k} p_i} \quad \text{(formula 3.3)}$$

unde p_i reprezintă ponderea scorului, iar x_i reprezintă scorul

III.2.2 Mediana

Mediana este o altă măsură a tendinței centrale şi **reprezintă valoarea care împarte şirul de măsurători în două parți egale; jumătate din şirul de date vor avea valori mai mici decât mediana în timp ce cealaltă jumătate vor avea valori mai mari decât mediana.** Să considerăm următorul şir de date:

Scoruri: 12, 14, 10, 11, 15, 14, 12, 20, 19, 20, 13, 11, 14, 8

Pentru a calcula mediana, primul pas este acela de a ordona crescător sau descrescător aceste date. Ordonând crescător şirul de mai sus, obţinem:

Scoruri: 8, 10, 11, 11, 12, 12, 13, 14, 14, 14, 15, 19, 20, 20

După definiţia medianei, în cazul nostru avem 14 valori. Prin urmare, mediana va fi valoarea care împarte acest şir ordonat în două părţi egale. Fiind 14 valori, mediana este situată la limita primelor şapte valori. Deoarece şirul este un **şir par**, mediana se situează, în cazul nostru, între valoarea 13 şi valoarea 14, mai precis la valoarea 13,5. În cazul unui **şir impar**, mediana este valoarea de la mijlocul unui şir. Dacă reluăm exemplul anterior, şi mai adăugăm un scor, obţinem:

Scoruri: 12, 14, 10, 11, 15, 14, 12, 20, 19, 20, 13, 11, 14, 8, 9

Şirul are acum 15 valori. Ordonând şirul, obţinem:

*Scoruri: 8, 9, 10, 11, 11, 12, 12, **13**, 14, 14, 14, 15, 19, 20, 20*

De data aceasta, la mijlocul acestui şir găsim valoarea 13, valoarea medianei.

Ordonând un şir, putem preciza poziţia fiecărui element în cadrul acelui şir. Astfel, dacă luăm şirul ordonat de 15 valori din exemplul de mai sus, obţinem:

Scoruri:	*8, 9, 10, 11, 11, 12, 12, **13**, 14, 14, 14, 15, 19, 20, 20*
Poziţia:	*1, 2, 3, 4, 5, 6, 7, **8**, 9, 10, 11, 12, 13, 14, 15*

Altfel spus, mediana nu este altceva decât poziţia rangului din mijloc în şirul ordonat de date. Într-o serie de la 1 la 15, poziţia din mijlocul şirului este evident, poziţia 8. Acesta este de altfel şi locul în care găsim mediana. Am accentuat asupra poziţiei în şir şi nu am discutat despre rang. În cazul în care avem două sau mai multe scoruri identice (în exemplul nostru fiind vorba despre 11, 12, 14 şi 20), poziţia în cadrul şirului exprimă poziţia fizică a

acelui element. Observăm că scorul 12 ocupă poziția 6 şi 7 în cadrul şirului, elementul 14 poziția 9, 10 şi 11 şi aşa mai departe. Când vorbim de rangul unui scor şi avem mai multe scoruri care se repetă, **rangul acestora va fi media aritmetică a poziţiilor pe care scorurile le ocupă în cadrul şirului.** Astfel, dacă reluăm exemplul nostru, îl putem completa şi cu rangul scorurilor, astfel:

Scoruri:	8,	9,	10,	11,	11,	12,	12,	13,	14,	14,	14,	15,	19,	20,	20
Poziția:	1,	2,	3,	4,	5,	6,	7,	8,	9,	10,	11,	12,	13,	14,	15
Rang:	1,	2,	3,	4.5,	4.5,	6.5,	6.5,	8,	10,	10,	10,	12,	13,	14.5,	14.5

Iată deci că scorul 11 apare pe poziția 4 şi 5 de două ori. Prin urmare, rangul acestor scoruri este media dintre 4 şi 5, adică 4,5. La fel, scorul 12 apare pe poziţiile 6 şi 7, rangul acestora fiind 6,5. Scorul 14 apare pe poziţiile 9, 10 şi 11, rangul acestora fiind 10 şi aşa mai departe.

Mediana poate fi calculată şi atunci când avem datele grupate[6]. Iată, am reluat exemplul anterior de calcul al mediei. Ceea ce trebuie să ştim, sunt frecvenţele cumulate şi numărul total de cazuri. În exemplul nostru, numărul total de cazuri este 30. Mediana fiind valoarea ce împarte grupul de subiecţi în două părţi egale, poate fi calculată după formula alăturată, unde *l* este limita inferioară a intervalului care conţine mediana, *i* este intervalul de clasă care conţine mediana, Poz_{Me} poziția medianei, f_c reprezintă frecvența cumulată a intervalului pre-median (totalul frecvenţelor situate înainte de mediană iar *f* reprezintă frecvența absolută a intervalului medianei.

$$Me = l + i\left(\frac{Poz_{Me} - fc_{Me-1}}{f}\right) \quad \text{(formula 3.4)}$$

[6] Asemenea mediei, „calculul" medianei prin acest procedeu reprezintă o aproximare a acesteia şi nu valoarea ei exactă

În cazul nostru, având 30 de măsurători, mediana va fi undeva între poziția 15 şi poziția 16, mai precis la 15,5 aşa cum am prezentat anterior. Practic, măsurătoarea care reprezintă poziția medianei este determinată de formula $Poz_{Me} = \frac{n+1}{2}$ dacă numărul de cazuri *n* este un număr par. Dacă *n* este un număr impar, poziția medianei este dată de formula $Poz_{Me} = \frac{n}{2}$. În cazul nostru, poziția medianei va fi la a $\frac{30+1}{2} = 15{,}5$-a măsurătoare. Dacă privim în tabelul de mai sus, la frecvențe cumulate, observăm că această poziție corespunde intervalului 179-182, intervalul imediat superior celei de a 14-a măsurători. Limita inferioară a acestui interval (*l)* este 179, intervalul de clasă (*i*) este 4 (179, 180, 181, 182), frecvența cumulată a intervalului pre-median (f_c) este 14 iar frecvența absolută (*f*) este 6. Avem acum toate datele necesare pentru a le înlocui în formulă şi obținem valoarea 180. Iată şi valoarea medianei, pe care o putem obține în cazul în care valorile sunt grupate în clase.

Tabel 3.5 – Distribuția statistică calcul mediană

Clasa	Distribuția în linii	f	f_c
171 – 174	|||||||||	9	9
175 – 178	|||||	5	14
179 – 182	||||||	6	20
183 – 186	||||	4	24
187 – 190	|||	3	27
191 -193	|||	3	30
		n=30	

$$Me = 179 + 4\left(\frac{\frac{30+1}{2} - 14}{6}\right) = 179 + 4\frac{15{,}5-14}{6} = 179 + 4\frac{1{,}5}{6} = 180$$

Mediana şi rangurile sunt intens folosite atunci când lucrăm cu date neparametrice sau atunci când media nu este un indicator fidel al tendinței centrale.

III.2.3 Modul

Doamnele, domnişoarele, ne pot spune cel mai bine ce este modul, căruia i se mai spune mod, modă sau valoare modală. Deci, doamnelor, ce este în definitiv moda? Ce înțelegem când spunem că ceva este la modă? Desigur, ceva este la modă atunci când o mare parte dintre oameni poartă, utilizează acel lucru. Spunem că cizmele roz sunt la modă, adică, foarte multe fete poar-

tă acele cizme roz. Iată un nou cuvânt comun derivat din statistică. **Prin urmare, modul nu este altceva decât categoria cu frecvența cea mai mare.** Foarte complicat, într-adevăr… Dacă avem şirul de date:

2, 15, 16, 18, 18, 21, 21, 21, 21, 25, 26

În acest caz observăm că există un singur 2, un singur 15, un singur 16, doi de 18, patru de 21, un 25 şi un 26. Care e categoria cu frecvența cea mai mare? Evident 21. Deci, modul este 21.

Modul poate fi calculat şi în cazul în care datele sunt grupate în clase[7]. Formula de calcul, în acest caz, este prezentată mai jos, în care l reprezintă limita inferioară a intervalului modal (intervalul cu frecvența absolută cea mai mare), f_m reprezintă frecvența absolută a intervalului modal, f_{m-1} frecvența absolută a intervalului pre-modal (intervalul aflat înaintea intervalului modal), f_{m+1} frecvența absolută a intervalului post-modal (intervalul aflat după intervalul modal) iar i este intervalul de clasă al clasei modale.

$$Mo = l + i\frac{f_m - f_{m-1}}{2f_m - f_{m-1} - f_{m+1}} \quad \text{(formula 3.5)}$$

În cazul nostru, intervalul modal este intervalul cu frecvența absolută cea mai mare, adică primul interval, 171-174 care are frecvența absolută 9. Intervalul pre-modal nu există, deci frecvența lui este 0 iar intervalul post-modal este intervalul 175-178 care are frecvența absolută 5. Evident, intervalul de clasă (i) este din nou 4 (171, 172, 173, 174). Aplicăm acum formula şi obținem 174.

Tabel 3.6 – Distribuția statistică calcul mod

Clasa	Distribuția în linii	f
171 – 174	|||||||||	9
175 – 178	|||||	5
179 – 182	||||||	6
183 – 186	||||	4
187 – 190	|||	3
191 -193	|||	3
		n=30

[7] Asemenea mediei şi medianei, „calculul" modului prin acest procedeu reprezintă o aproximare a acestuia şi nu valoarea sa exactă

$$Mo = 171 + 4\frac{9-0}{2\times 9-0-5} = 171 + 4\frac{9}{13} = 171 + \frac{40}{13} = 174$$

III.2.4 Precizia indicatorilor tendinței centrale

Am întâlnit, iată, trei indicatori ai tendinței centrale. Care dintre cei trei indicatori caracterizează mai bine setul nostru de date? Care este mai precis? Media veți spune toți şi poate aşa şi este. Dar haideți să ne mai gândim… ce înseamnă totuşi tendința centrală? Evident un indicator al scorurilor tipice din setul nostru de date, un model statistic al unei realități. Cu alte cuvinte, un indicator care descrie cât mai exact setul de date şi care ne poate spune cum se prezintă majoritatea datelor. Haideți să considerăm următorul set de date:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

În acest caz, media este 5,5, iar mediana tot 5,5. Să modificăm puțin ultima cifră din setul nostru de date:

1, 2, 3, 4, 5, 6, 7, 8, 9, 20

Iată că, media devine acum 6,5 în timp ce mediana rămâne 5,5. Haideți să modificăm mai mult ultima cifră:

1, 2, 3, 4, 5, 6, 7, 8, 9, 100

Acum media va fi 14,5 în timp ce mediana rămâne tot 5,5. Puteți intui care indicator caracterizează cel mai bine tendința centrală în ultimul set de date? Mediana veți spune şi aveți dreptate. Iată că, **atunci când avem scoruri extreme, foarte mari sau foarte mici în comparație cu setul de date, mediana este un indicator mai bun decât media.** Despre modul în care depistăm scorurile extreme vom discuta într-un alt capitol. Pentru a înțelege mai bine, luăm ca exemplu veniturile unor persoane, după cum urmează, în RON:

100, 105, 110, 115, 120, 125, 130, 135, 140

În cazul nostru, media este 120, iar mediana este tot 120. Putem concluziona ca aceşti oameni sunt relativ săraci. Dar haideți să modificăm ultima valoare, în loc de 140 vom stabili 1400. Şirul devine:

100, 105, 110, 115, 120, 125, 130, 135, 1400

În acest caz media devine 260 în timp ce mediana rămâne tot 120. Putem oare afirma că aceşti nouă oameni câştigă în medie 260 RON? Categoric, nu, observăm că în afară de ultima persoană, nici una dintre cele opt rămase nu câştigă nici măcar 150 RON, darămite 260. Dacă ne-am lua după medie, am putea spune că oamenii aceştia sunt relativ bogați, ceea ce este complet fals. Iată că acest scor extrem face media nereprezentativă pentru setul de date. Cel mai bun indicator rămâne tot mediana. Desigur, mediana nu este un indicator atât de sensibil ca şi media, însă în momentul în care nu putem folosi media datorită valorilor extreme, folosim mediana ca indicator al tendinței centrale.

O altă observație o facem în situația datelor neparametrice aflate la un nivel nominal de măsurare. Spre exemplu, urmărim culoarea ochilor unor elevi dintr-o clasă şi obținem:

Albaştri = 5, Verzi = 6, Căprui = 15, Negri = 10

În acest caz nu putem folosi media ca indicator al tendinței centrale pentru că nu are sens (care e media? Un soi de albastro-verzuio-căprui-negri????) dar nici mediana pentru că nu putem ierarhiza categoriile. Nu putem spune care culoare de ochi e superioară căreia. Singurul indicator al tendinței centrale rămâne modul. Bazându-se pe o simplă numărare de frecvențe, are sens să vorbim aici de mod şi să afirmăm că modul este „căprui", aceasta fiind categoria cu frecvența cea mai mare.

Iată deci că **pentru date neparametrice aflate la un nivel nominal de măsurare, singurul indicator al tendinței centrale ce poate fi calculat este modul.**

Modul poate fi de asemenea un bun indicator al tendinței centrale şi în cazul datelor parametrice. Să considerăm următorul exemplu:

1, 2, 2, 2, 2, 2, 2, 2, 3, 4, 5, 6, 7, 8
1, 2, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

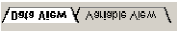
În primul caz observăm că avem o frecvență foarte mare a categoriei „2" comparativ cu al doilea caz. Deci, în primul caz modul ar fi cel mai bun indicator al tendinței centrale (al scorurilor tipice din serie) Şi în al doilea caz modul este tot doi, însă observăm că apare doar o singură dată în plus în comparație cu celelalte valori spre deosebire de primul caz în care apare de şase ori în plus. Iată că în al doilea caz modul nu este un bun indicator al tendinței centrale şi se preferă media, în timp ce în primul caz se preferă modul, acesta indicând cel mai bine scorurile tipice.

În situația în care frecvența unei categorii este mult mai mare în comparație cu frecvența celorlalte categorii, cel mai bun indicator al tendinței centrale este modul.

Cel mai sensibil indicator este media deoarece se bazează pe scoruri, însă media induce erori dacă avem scoruri extreme, mult mai mari sau mult mai mici în comparație cu celelalte scoruri. Mediana se bazează pe ranguri şi de aceea este mai puțin precisă decât media însă şi insensibilă la aceste scoruri extreme. Modul se bazează pe frecvența de apariție şi este cel mai imprecis indicator al tendinței centrale dar şi cel mai flexibil. Îl putem folosi în cazul datelor categoriale sau în cazul în care o categorie are o frecvență mult mai mare în comparație cu celelalte categorii. De asemenea, pentru date categoriale (neparametrice la nivel nominal de măsurare), putem folosi doar modul ca indicator al tendinței centrale. Pentru date ordinale, unde putem stabili ierarhii, putem folosi modul sau mediana iar pentru date parametrice (scalare) putem folosi modul, mediana sau media în funcție de tipul datelor.

III.2.5 Obţinerea indicatorilor tendinţei centrale în SPSS

Vă mai amintiţi, cred, de baza de date creată în SPSS în capitolul unu şi pe care am numit-o IQ. Haideţi să o deschidem. Dacă nu reuşiţi, revedeţi primul capitol.

Acum vom introduce câteva date în SPSS. Trecem deci în modul **„Data View”** prin apăsarea tabulatorului din partea din stânga jos a programului SPSS. 

6 :

	nume	varsta	sexul	scoala	iq
1	Agapie Ion	35	1	3	100
2	Agapie Manuela	20	2	3	120
3	Ionescu Carme	21	2	3	124
4	Dominte Vasile	24	1	2	121
5	Popescu Ioana	29	2	3	124
6	Pop Gheorghe	28	1	3	101
7	Diaconu Dana	29	2	4	110
8	Alexandru Virgil	29	1	4	124
9	Sandu Maria	24	2	5	121
10	Sandulache Georgeta	31	2	5	121
11	Dancu Dan	26	1	5	100
12	Vasiloiu Ioan	20	1	5	101
13	Dumitrescu Paraschiv	38	1	5	105
14	Olteanu Costica	37	1	1	108
15	Popa Dorina	38	2	1	104
16	Vasile Doina	29	2	2	110
17	Zancea Elena	24	2	3	124
18	Dogaru Emil	26	1	3	111
19	Panciuc Daniel	29	1	3	105
20	Ogaru Sandu	26	1	3	108
21	Oleg Felicia	24	2	4	107
22	Dorneanu Sanda	37	2	4	109
23	Barbu Ion	30	1	4	110
24	Cozan Valeria	35	2	4	108
25	Trofin Alexandru	24	1	3	124
26	Barbinte Floarea	32	2	3	121
27	Kelt Iosef	31	1	3	104
28	Jipa Maria	34	2	4	105
29	Jipa Andrei	39	1	5	103
30	Marinescu Cristian	31	1	5	107
31					

Figura 3.1 – Baza de date populată cu date – afişarea asociaţiilor sub formă de coduri

Se afişează o foaie de lucru, un tabel similar tabelului din Excel, în care pe prima coloană, cea de culoare gri (capul de tabel), întâlnim numele variabilelor definite anterior.

Modul de lucru este similar celui din Excel. Se scriu pur şi simplu datele în celule. Atenţie! Avem două variabile pentru care am asociat valori: variabila „sexul” şi variabila „scoala”. Aici nu vom scrie cuvântul „Masculin” sau „Feminin” ci vom tasta 1 pentru „Masculin” şi 2 pentru „Feminin”. La fel şi în cazul

variabilei „scoala” vom tasta 1 pentru „Primare”, 2 pentru „Gimnaziu” şi aşa mai departe.

Baza dumneavoastră de date trebuie să arate în final ca în figura de mai jos. O mică precizare. Pe bara de instrumente, penultimul buton se numeşte „Value Labels” şi arată ca în imaginea alăturată. La o apăsare a acestuia, butonul se activează iar în baza de date se afişează etichetele asociaţilor făcute, ca în figura alăturată. La o nouă apăsare etichetele asociaţiilor nu se mai afişează ci se afişează valorile asociaţiilor ca în figura mare de mai sus.

nume	varsta	sexul	scoala	iq
Agapie Ion	35	Masculin	Liceu	100
Agapie Manuela	20	Feminin	Liceu	120
Ionescu Carme	21	Feminin	Liceu	124
Dominte Vasile	24	Masculin	Gimnaziu	121
Popescu Ioana	29	Feminin	Liceu	124
Pop Gheorghe	28	Masculin	Liceu	101
Diaconu Dana	29	Feminin	Postliceala	110
Alexandru Virgil	29	Masculin	Postliceala	124
Sandu Maria	24	Feminin	Facultate	121

Figura 3.2 – Baza de date cu afişarea etichetelor asocierilor

După ce am introdus cele 30 de date, vom salva baza de date şi vom iniţia procedura de extragere a indicatorilor tendinţei centrale. Indicatorii tendinţei centrale ne sunt furnizaţi prin intermediul numeroaselor proceduri statistice. Cele mai simple comenzi, pentru a nu complica inutil expunerea, le găsim în meniul **„Analyze”** şi apoi meniul **„Descriptive Statistics”**. În traducere, analize – statistici descriptive. Primele trei comenzi din acest submeniu care se deschide, pot extrage indicatorii tendinţei centrale. Vom analiza doar prima comandă, comanda **„Frequencies…”** urmând ca celelalte două să le discutăm cu alte ocazii deoarece presupun şi alte informaţii.

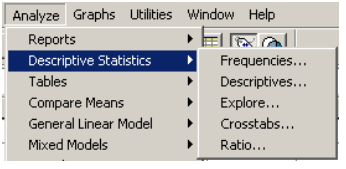


Figura 3.3 – Lansarea procedurilor de analiză descriptivă pe frecvenţe

Comanda **„Frequencies…”** sau frecvenţe prezintă cele trei puncte, fapt care va duce la deschiderea unei noi casete de dialog. Daţi clic pe această comandă şi veţi obţine o casetă ca cea din imaginea de mai jos.

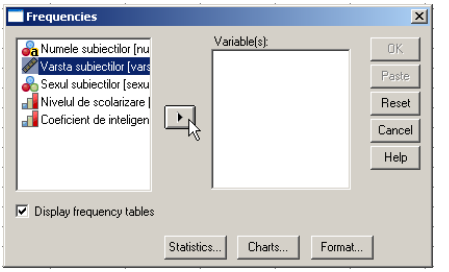


Figura 3.4 – Caseta de analize descriptive utilizându-se frecvenţele

În partea stângă se află o listă cu toate variabilele din baza dumneavoastră de date. Alături de această listă se află un buton de transfer (acel buton cu o săgeată pe el). Urmează apoi lista **„Variable(s)”** care înseamnă variabilele pe care le introducem spre analiză. Pentru a introduce o variabilă spre analiză trebuie să o selectăm, printr-un clic simplu pe numele acesteia, din lista de variabile din partea stângă. O variabilă selectată are fundalul albastru (sau o altă culoare în funcţie de schema de culori a calculatorului dumneavoastră). După ce o selectaţi, apăsaţi butonul de transfer (care va avea săgeata orientată spre lista **„Variable(s)”** pentru a transfera variabila dumneavoastră în lista variabilelor ce urmează a fi analizate. În figura de mai sus, transferăm variabila „Varsta subiectilor” din lista variabilelor din baza de date în lista variabilelor supuse analizei. În final va trebui să obţineţi o imagine ca cea de mai sus. Observaţi că dacă am transferat variabila „Varsta subiectilor” în lista variabilelor supuse analizei, butonul de transfer şi-a schimbat sensul. Acum dacă apăsăm pe el „scoatem” variabila noastră din lista variabilelor supuse analizei (după ce în prealabil, evident, o selectăm) şi o introducem în lista variabilelor totale din baza de date. Oricum, acest buton de transfer ne va însoţi permanent în analizele noastre deoarece va trebui să „spunem” programului ce variabile să analizeze şi pe care să nu le analizeze.

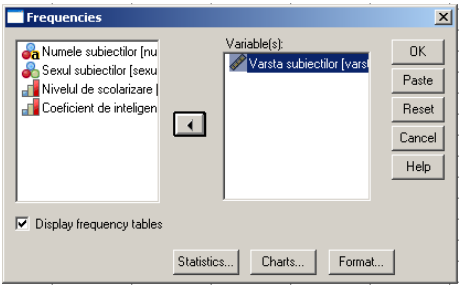


Figura 3.5 – Includerea variabilelor spre analiză

Vom trece în revistă acum toate comenzile din această fereastră cu precizarea că unele dintre ele vor fi abordate în detaliu în cadrul altor capitole.

Caseta de bifare **„Display frequency tables”** informează aplicația SPSS dacă să afişeze tabelul frecvențelor (în condițiile în care această casetă o bifați) sau să nu-l afişeze (în condițiile în care nu o bifați). Tabelul frecvențelor este acel tabel general, studiat la capitolul despre organizarea datelor, înainte de a grupa valorile în clase. Iată că acum am ales ca acest tabel al frecvențelor să fie afişat.

În partea dreaptă a acestui formular se află un număr de cinci butoane de acționare.

Butonul **„OK”** este butonul pe care vom apăsa pentru a lansa analiza, după ce am configurat modul în care dorim ca analiza să fie efectuată.

Butonul **„Paste”** este un buton pe care nu îl vom folosi deloc. El permite „lipirea” codului care efectuează analiza curentă într-o fereastră de sintaxă de comenzi. El se adresează utilizatorilor care doresc să facă programare în SPSS utilizând limbajul propriu de programare al acestei aplicații.

Butonul **„Reset”** permite reinițializarea ferestrei la starea ei originală. Cu alte cuvinte, toate variabilele din lista de analiză sunt eliminate, toate configurările de analiză pe care le-am făcut sunt inițializate la valoarea lor implicită. Este un buton pe care îl vom folosi destul de frecvent.

Butonul **„Cancel”** este butonul prin care renunțăm la analiză şi închidem această fereastră fără a mai efectua analiza.

Butonul **„Help”** este butonul cu ajutorul căruia obținem asistență şi informații (în limba engleză) referitoare la funcționarea acestei proceduri. Dacă ştiți engleză, nu ezitați să-l utilizați. SPSS are un tutorial excelent pe care îl puteți parcurge şi înțelege cu succes.

În partea de jos a formularului se află trei butoane care permit configurarea statisticilor pe care le vom calcula.

Butonul **„Statistics…”** este butonul pe care vom şi apăsa acum pentru a calcula indicatorii tendinţei centrale. După cum îi spune şi numele, cu ajutorul acestui buton putem stabili ce statistici descriptive vom analiza. Iată că cele trei puncte de suspensie deschid acum o nouă fereastră. Fereastra are mult mai multe opţiuni. Pentru moment ne vom rezuma la cele care ne interesează. Observăm că în partea dreaptă a acestei noi ferestre găsim secţiunea **„Central Tendency”** adică exact ce ne interesează pe noi. În această secţiune observăm casetele de bifare **„Mean”** (media), **„Median”** (mediana) şi **„Mode”** (modul). Iată elementele pe care dorim să le calculăm. Vom bifa deci, una câte una, aceste casete prin efectuarea unui clic simplu pe ele astfel încât formularul să arate ca în figura de mai sus. Mai există şi caseta de bifare **„Sum”** (suma) prin bifarea căreia obţinem suma valorilor din variabila respectivă. Variabila noastră este „vârsta subiecţilor” şi nu ne interesează să însumăm vârstele celor 30 de subiecţi. Ar fi un indicator fără nici o relevanţă pentru noi.

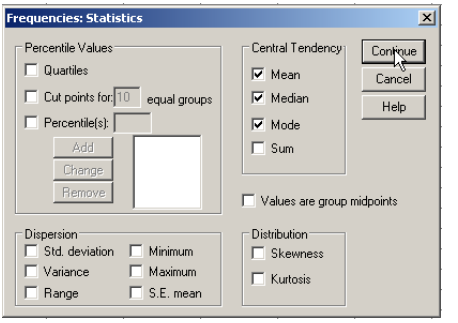


Figura 3.6 – Fereastra de alegere a indicatorilor tendinţei centrale

După ce am bifat cele trei casete, apăsăm butonul **„Continue”** (continuare) pentru a părăsi această fereastră şi a ne întoarce în cea iniţială. Acum am comunicat programului SPSS ce indicatori statistici să extragă.

Butonul **„Chart…”** şi **„Format…”** permite afişarea unor grafice, respectiv configurarea modalităţi în care vor fi afişate rezultatele. Despre aceste butoane vom discuta cu altă ocazie.

Cam asta ar fi tot. Practic am spus programului SPSS să calculeze media, mediana şi modul pentru variabila „Varsta subiectilor” cu afişarea

tabelului de frecvențe. Nu ne rămâne acum decât să apăsăm butonul **„OK”** și să vedem ce se întâmplă.

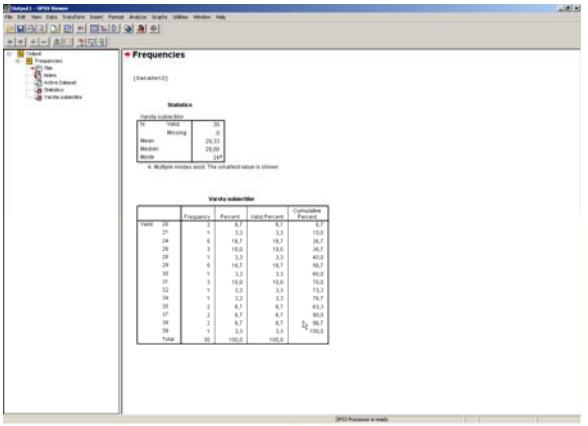


Figura 3.7 – Fereastra de afișare a rezultatelor

Iată că a apărut fereastra de rezultate sau fereastra de **„Output”** Nu vom intra acum în detalii explicative ale acestei ferestre pentru că s-ar putea să vă speriați de prea multe informații și să treceți la „Solitaire”. Ne vom concentra asupra celor două tabele pe care le-a generat SPSS în partea din dreapta a ferestrei de rezultate (partea cea mare) care nu reprezintă altceva decât cele două tabele pe care le-am solicitat. Primul tabel conține indicatorii tendinței centrale pe care i-am configurat să fie calculați în etapa anterioară.

Tabel 3.7 – Indicatorii tendinței centrale

Statistics

Varsta subiectilor

N	Valid	30
	Missing	0
Mean		29,33
Median		29,00
Mode		24[a]

a. Multiple modes exist. The smallest value is shown

Tabelul se numește **„Statistics”** și conține în partea de sus numele variabilei, „Varsta subiectilor”, pentru a putea identifica variabila cu care operăm. Apoi ne dă numărul de cazuri, **„N”** și ne spune că sunt 30 de date valide (**„Valid”**) și zero date lipsă (**„Missing”**) Cu alte cuvinte toți cei 30 de subiecți au înregistrată vârsta lor. Nu există nici un subiect la care să fi uitat să introducem această valoare. În continuare ne spune că media de vârstă este 29,33 (**„Mean”**), mediana este 29 (**„Median”**) iar modul…. avem o problemă. SPSS ne informează în josul tabelului că sunt mai multe valori cu frec-

venţa cea mai mare şi o va afişa pe cea mai mică, 24 (**„Multiple modes exist. The smallest value is shown”**) Acesta este deci un singur mod real, cel inferior şi urmează să vedem care este şi cealaltă valoare modală deoarece avem o distribuţie cu mai multe moduri (multimodală).

Următorul tabel este tabelul de frecvenţe. Şi aici se afişează numele variabilei. Probabil că acest tabel vă pare extrem de cunoscut. Desigur, nu avem diagrama în linii însă tabelul seamănă frapant cu cel descris de noi la capitolul despre sistematizarea datelor. Iată că acum toate valorile sunt ordonate crescător şi avem frecvenţa absolută (**„Frequency”**) pentru fiecare valoare, frecvenţa relativă (**„Percent”**) şi frecvenţa relativă cumulată (**„Cumulative percent”**). Mai apare o coloană, frecvenţa relativă validă (**„Valid percent”**) care este identică cu frecvenţa relativă deoarece nu sunt valori lipsă în şirul nostru de date. Iată că acum putem să găsim şi celălalt mod. Ştim că primul mod este 24 iar acesta este modulul inferior. Ne uităm în tabelul de frecvenţe şi vedem că valoarea 24 are frecvenţa 5. Căutăm acum o valoare mai mare de 24 cu aceeaşi frecvenţă şi observăm că este 29. Iată că cele două valori modale sunt 25 şi 29.

Tabel 3.8 – Distribuţia statistică

Varsta subiectilor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	2	6,7	6,7	6,7
	21	1	3,3	3,3	10,0
	24	5	16,7	16,7	26,7
	26	3	10,0	10,0	36,7
	28	1	3,3	3,3	40,0
	29	5	16,7	16,7	56,7
	30	1	3,3	3,3	60,0
	31	3	10,0	10,0	70,0
	32	1	3,3	3,3	73,3
	34	1	3,3	3,3	76,7
	35	2	6,7	6,7	83,3
	37	2	6,7	6,7	90,0
	38	2	6,7	6,7	96,7
	39	1	3,3	3,3	100,0
	Total	30	100,0	100,0	

Aceasta este metoda cea mai simplă şi cea mai completă de extragere a indicatorilor tendinţei centrale. Închidem acum fereastra de rezultate prin efectuarea unui clic pe butonul *„X”* din colţul din dreapta sus al acesteia.

Concluzii:

- *Populația reprezintă totalitatea unor elemente dintr-un anumit spațiu la un anumit moment dat;*
- *În mod curent nu putem studia populația şi de aceea lucrăm cu eşantioane extrase din populație care să fie reprezentative, adică să respecte caracteristicile populației pe anumite criterii;*
- *Cele mai utilizate metode de eşantionare sunt:*
  - *Eşantionarea simplu randomizată;*
  - *Eşantionarea pe cote;*
  - *Eşantionarea pe straturi;*
  - *Eşantionarea pe clusteri*
- *Indicatorii tendinței centrale sunt media, mediana şi modul;*
- *Media este cel mai precis indicator, urmat fiind de mediană şi de mod;*
- *Pentru date categoriale singurul indicator ce poate fi utilizat este modul;*
- *Pentru date ordinale se poate utiliza modul sau mediana;*
- *Pentru date parametrice se foloseşte modul, mediana sau media;*
- *Media nu se foloseşte atunci când avem scoruri extreme, foarte mici sau foarte mari în distribuția noastră;*
- *Modul se foloseşte la date parametrice atunci când frecvența unei valori este foarte mare în comparație cu frecvența celorlalte valori;*
- *Media nu poate fi folosită la date neparametrice*

Întrebări şi exerciții:

- *Calculați media următoarei distribuții statistice precizând etapele de rezolvare:*
- *Calculați mediana şi rangurile distribuției statistice a înălțimii subiecților din exemplul din capitolul doi.*
- *Precizați care este modul distribuției statistice a înălțimii subiecților din exemplul din capitolul anterior? Sunt cumva mai multe valori?*
- *Calculați media, mediana şi modul în cazul următoarelor seturi de date:*
  - *12, 23, 9, 6, 14, 14, 12, 25, 9, 12*
  - *1, 4, 5, 6, 19, 1, 5, 3, 16, 12, 5, 4*
  - *32, 56, 91, 16,32, 5, 14,62,19, 12*
- *Care este cel mai adecvat indicator al tendinței centrale pentru fiecare dintre următoarele seturi de date:*
  - *1, 23, 25, 26, 27, 23, 29, 30*
  - *1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3 ,3 ,4, 5*
  - *1, 1, 2, 3, 4, 1, 2, 6, 5, 8, 3, 4, 5, 6, 7*
  - *1, 101, 104, 106, 111, 108, 109, 200*

Clasa	**f**
3 – 5	3
6 – 8	5
9 – 11	9
12 – 14	16
15 – 17	10
18 – 20	4
21 – 23	3
24 – 26	1

IV. MEDIA POPULAȚIEI. REPREZENTĂRI GRAFICE ALE TENDINȚEI CENTRALE

În acest capitol se va discuta despre:

- Teorema limitei centrale;
- Reprezentări grafice ale datelor;
- Interpretări ale reprezentărilor grafice de date;

După parcurgerea capitolului, cursanții vor fi capabili să:

- Înțeleagă termenul de medie a eşantioanelor;
- Reprezinte grafic datele utilizând SPSS;
- Interpreteze semnificația reprezentărilor grafice.

Atunci când lucrăm cu un eşantion, deci cu un număr relativ redus de subiecți, calculul mediei nu este o operațiune atât de dificilă. Indiferent dacă folosim formula de definiție a mediei sau o altă modalitate de calcul, în final vom reuşi obținerea acestui indicator al tendinței centrale. Ce ne facem însă atunci când avem de stabilit un asemenea indicator la nivelul populației? Dacă ne propunem să aflăm media înălțimii populației din România, cum am putea oare proceda? Teoretic ar trebui să măsurăm înălțimea fiecărui om din România şi apoi să împărțim suma acestor înălțimi la populația totală a țării. Practic acest lucru este imposibil. Avem oare vreo variantă prin care să putem afla totuşi acest indicator? Veți afirma că putem extrage un eşantion din populație şi aveți dreptate. Vom extrage un eşantion, însă există posibilitatea să selectăm în acest eşantion oameni mai curând înalți. Dacă, de exemplu, obținem o medie a înălțimii persoanelor din eşantion de 178 centimetri putem afirma că aceasta este media populației din România? Nu cred că putem face această supoziție. Putem însă să mai extragem un eşantion şi obținem o medie a înălțimii populației de 165 centimetri. Această nouă valoare este evident diferită de prima evaluare şi putem suspecta că am ales în eşantion persoane mai curând scunde. Prin urmare, care dintre cele două medii se apropie mai mult de media înălțimii populației din România? Răspunsul este că nici una.

Însă, vom face media acestor două eşantioane. Deci, adunăm 178 cu 165 şi împărţim la doi. Obţinem 171,5 centimetri, un indicator mai precis pentru media populaţiei. Iată deci că media eşantioanelor este un indicator care aproximează mai bine media unei populaţii. Cu cât avem mai multe eşantioane cu atât putem aproxima mai precis media populaţiei. Această caracteristică a mediei eşantioanelor de a tinde către media populaţiei poartă numele în statistică de **„teorema limitei centrale”**. Asupra acestui aspect vom reveni însă şi vom avea ocazia să îl tratăm amănunţit într-un capitol ulterior.

IV.1 Reprezentări grafice

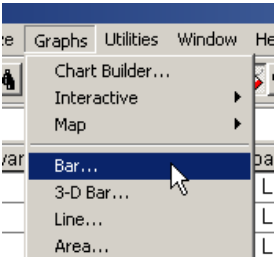


Figura 4.1 – Meniul pentru graficul cu bare

Am studiat în primul capitol o serie de reprezentări grafice uzuale. Vom vedea cum putem reprezenta grafic datele folosind SPSS şi ce semnificaţie au aceste reprezentări. Aşadar, vom deschide baza de date creată de noi anterior, baza de date IQ, şi ne propunem să reprezentăm datele noastre printr-un grafic cu bare.

Vom folosi pentru aceasta meniul **„Graphs”** şi apoi opţiunea **„Bar…”** În general, în meniul „Graphs” găsim toate posibilităţile de reprezentare grafică a datelor pe care SPSS le pune la dispoziţie[8].

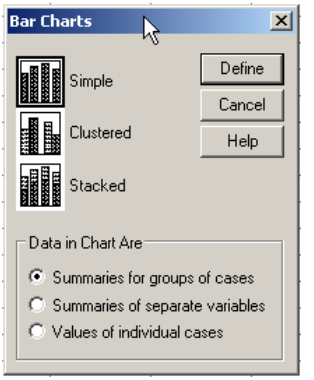


Figura 4.2 – Alegerea tipului de grafic cu bare

În momentul în care am accesat opţiunea „Bar...” apare o nouă fereastră în care va trebui să alegem tipul de grafic cu bare pe care dorim să-l reprezentăm. Avem la dispoziţie trei variante: **„Simple”** (simplu), **„Clustered”** (pe clusteri) şi

[8] În SPSS versiunea 15, modalitatea de construcţie a graficelor diferă radical, fiind mult mai simplă şi mai inteligentă. Vom aborda aspectele grafice ale aplicaţiei SPSS 15 la sfârşitul acestui volum şi în volumele viitoare

„Stacked” (aditiv). Graficul simplu ne permite reprezentarea cu bare doar a variabilei noastre (în cazul nostru vom folosi reprezentarea cu bare a coeficientului de inteligență, „iq”) Graficul pe clusteri îl folosim dacă dorim să includem o variabilă categorială (de exemplu dacă dorim să prezentăm comparativ coeficientul de inteligență al bărbaților şi al femeilor) iar graficul aditiv prezintă valorile una deasupra alteia şi se foloseşte atunci când urmărim să vedem cu cât creşte o variabilă în timp sau care este rezultatul cumulării mai multor variabile. În cazul nostru vom folosi graficul cu bare simplu. Aşadar vom efectua clic pe imaginea graficului „Simple” pentru a-l selecta.

Există, după cum puteți observa, un număr de trei opțiuni pe care le veți folosi: prima opțiune, **„Summaries for group of cases”** arată că SPSS va parcurge categoriile unei singure variabile şi va afişa, pentru fiecare bară, categoria corespunzătoare. A doua opțiune, **„Summaries of separate variables”** arată că două sau mai multe variabile vor fi parcurse iar barele vor arăta fiecare dintre variabile. În sfârşit, opțiunea **„Values of individual cases”** ne spune că SPSS va lucra cu o singură variabilă iar barele vor reprezenta valorile sale. Aceasta este de fapt şi opțiunea care ne interesează pe noi. Alegem deci această opțiune după care apăsăm butonul **„Define”** pentru a defini variabilele ce vor fi reprezentate.

Apare o nouă fereastră cu mai multe elemente. Recunoaştem în partea stângă lista variabilelor din baza noastră de date. În partea dreaptă, primul element îl reprezintă caseta **„Bars Represent”** care se referă la ceea ce vor reprezenta barele noastre. Cu alte cuvinte, ce variabilă va fi reprezentată prin bare. Variabila pe care dorim să o reprezentăm este „Coeficient de inteli-

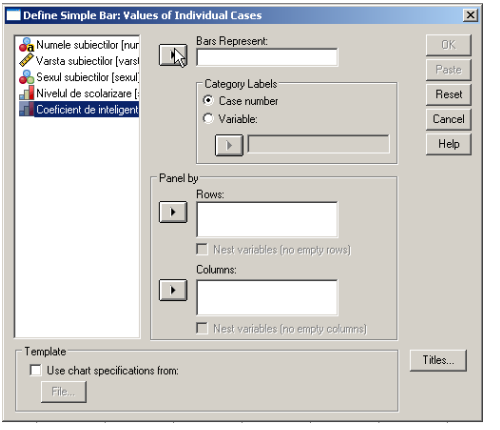


Figura 4.3 – Opțiunile de reprezentare ale graficului simplu cu bare

genţă". O vom selecta şi apoi vom apăsa butonul de transfer pentru a o include în caseta „Bars Represent" ca în figura de mai sus. Pentru moment atât avem de făcut în vederea trasării graficului cu bare. Există aici şi alte opţiuni pe care le putem folosi. Secţiunea **„Category Labels"** ne permite să alegem dacă pe axa categoriilor (axa O_X) va fi afişat numărul fiecărui caz (opţiunea **„Case number"**) sau valorile unei alte variabile categoriale pe care va trebui să o specificăm (opţiunea **„Variable:"**) Secţiunea **„Panel by"** permite gruparea barelor în funcţie de o altă variabilă categorială. Gruparea poate fi efectuată fie pe linii, fie pe coloane, cazuri în care va trebui să specificăm şi variabilele categoriale în funcţie de care se va face gruparea, introducându-le în listele **„Rows:"** sau **„Columns:"**. Dacă avem vreun şablon de grafic preferat, descărcat de pe Internet sau construit de noi, îl putem folosi pentru a desena graficul nostru bifând caseta **„Use chart specification from:"** şi apoi apăsând butonul **„File…"** care se activează şi ne invită să încărcăm acel fişier. Putem de asemenea da un titlu graficului nostru apăsând butonul **„Titles…".** Utilizarea acestui buton v-o lăsăm dumneavoastră ca exerciţiu. După ce am terminat de stabilit variabila pentru care dorim să trasăm graficul, apăsăm butonul **„OK"** care se activează. SPSS va ge-

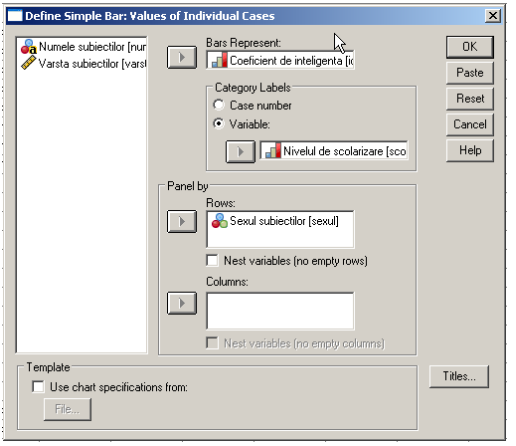


Figura 4.4 – Reprezentarea prin grafic cu bare a QI pe nivele se şcolarizare şi în funcţie de sex

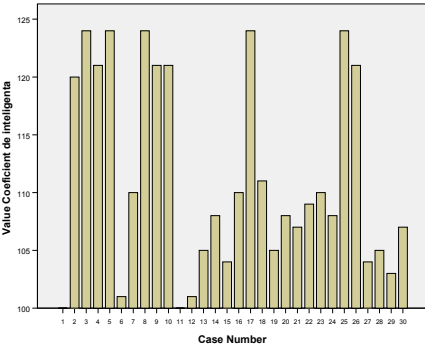


Figura 4.5 – Graficul cu bare simplu al coeficientului de inteligenţă

nera graficul în fereastra de rezultate (Output) rezultând un grafic ca în figura de mai sus. Iată aşadar reprezentarea cu bare care ne interesează.

Pentru a exemplifica şi celelalte facilităţi ale acestei reprezentări grafice, iată, în figura alăturată, un grafic cu bare în care am folosit ca variabilă etichetă „Nivelul de şcolarizare" iar graficul este împărţit pe linii după variabila „Sexul subiecţilor". Lansând procedura de desenare prin apăsarea butonului „OK", obţinem următorul grafic cu bare:

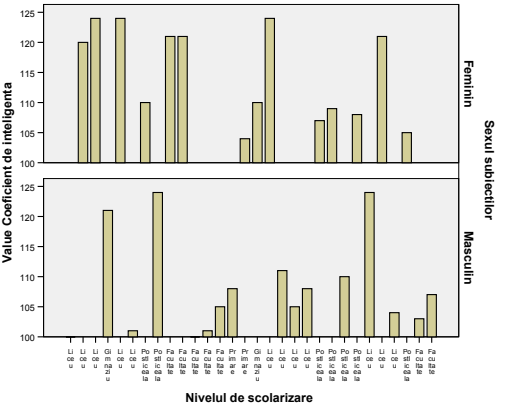


Figura 4.6 - Reprezentarea prin grafic cu bare a QI pe nivele se şcolarizare şi în funcţie de sex

Observăm două grafice, în cel de jos fiind reprezentat coeficientul de inteligenţă al bărbaţilor iar în cel de sus al femeilor. Dacă privim cu atenţie pe axa O_X, observăm că pentru fiecare bară este indicat şi nivelul de şcolarizare al subiectului respectiv. După ce veţi stăpâni la un nivel suficient de ridicat lucrul în SPSS, veţi putea face combinaţii între variabile pentru a realiza grafice intuitive cu care să vă susţineţi demonstraţiile statistice.

***Exerciţiu**:*

Reprezentaţi, folosind un grafic cu bare, valorile variabilei „Varsta subiectilor"

În primul capitol discutam de asemenea de poligonul frecvenţelor absolute, al valorilor individuale, al frecvenţelor relative şi al frecvenţelor cumulate absolute sau relative. Folosind SPSS vom putea desena poligonul valorilor individuale folosind un grafic cu linii. Vom accesa, pentru acesta, tot meniul **„Graphs"** şi apoi submeniul **„Line…"** La fel ca şi în graficul cu

bare, se afişează o nouă fereastră care ne invită să alegem tipul graficului pe care îl dorim. Avem de ales între un grafic simplu (**„Simple”**) care ne interesează şi pe noi, între un grafic cu linii multiple (**„Multiple”**) ce permite trasarea a două sau mai multe variabile pe acelaşi grafic şi un grafic de tip pivot (**„Drop-line”**) cu ajutorul căruia putem reprezenta două sau mai multe categorii de date. Pentru exemplul nostru, vom alege graficul de tip simplu. La fel ca şi în cazul graficului cu bare, şi aici avem cele trei opţiuni cu exact aceeaşi semnificaţie. Vom alege, ca şi în exemplul anterior, opţiunea **„Values of individual cases”** deoarece dorim să reprezentăm un poligon al frecvenţelor individuale. Pentru a putea lansa procedeul de definire a graficului, putem apăsa butonul **„Define”**.

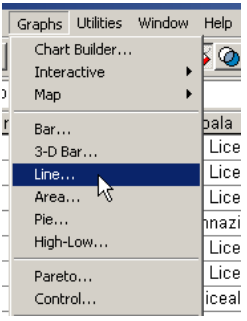


Figura 4.7 – Meniul pentru graficul cu linii

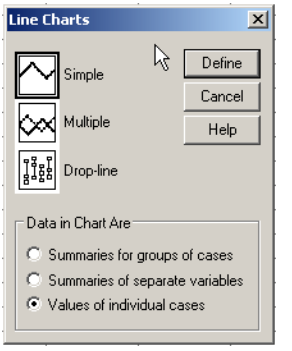


Figura 4.8 – Selectarea tipului de grafic cu linii

Observăm apariţia unei ferestre identice cu fereastra pe care am întâlnit-o la graficul cu bare şi care are exact aceeaşi funcţionalitate. Nu vom intra din nou în detalii asupra elementelor pe care le găsim aici, deoarece ele au fost discutate la graficul cu bare. Vom transfera doar variabila „Coeficient de inteligenţă” din lista variabilelor în lista **„Line Represents”** pentru a indica programului SPSS variabila pe care urmează să o reprezinte prin linii. De fapt, aceasta este singura diferenţă între cele două ferestre. În fereastra

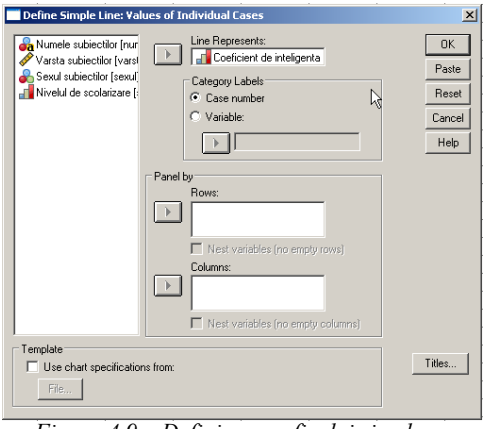


Figura 4.9 – Definirea graficului simplu cu linii

anterioară acest element se numea „Bar Represent" deoarece era vorba de un grafic cu bare iar aici se numeşte „Line Represents" deoarece suntem în faţa unui grafic cu linii. Nu ne rămâne acum decât să apăsăm butonul **„OK"** care vedem că s-a activat, pentru a lansa procedura de desenare a graficului. Iată, în fereastra de rezultate apare după câteva momente şi poligonul frecvenţelor individuale pe care l-am solicitat.

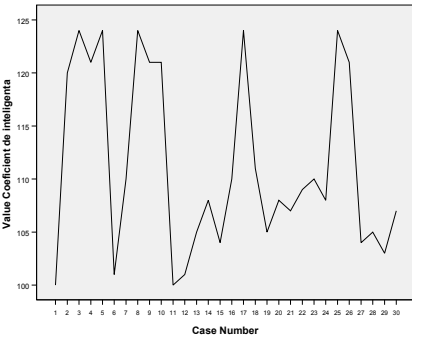


Figura 4.10 – Poligonul frecvenţelor absolute pentru coeficientul de inteligenţă

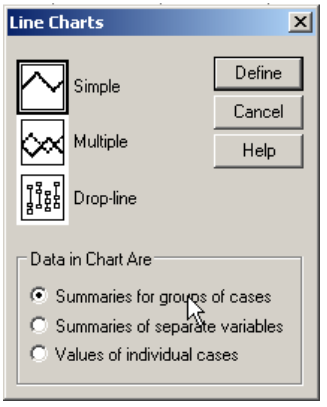


Figura 4.11 – Configurarea graficului pentru frecvenţele cumulate

Cum am putea totuşi să reprezentăm grafic poligonul frecvenţelor absolute cumulate? Ne amintim, din capitolul despre organizarea datelor, că frecvenţele cumulate se calculează prin adăugarea frecvenţei unei valori la frecvenţa cumulată a valorii anterioare, şirul valorilor fiind un şir ordonat. La fel, am văzut că la desenarea unui grafic cu linii (sau cu bare) în SPSS, avem la dispoziţie mai multe opţiuni iar noi nu am folosit decât opţiunea „Values of individual cases". De data aceasta vom merge tot în meniul **„Graphs",** opţiunea **„Line.. ".** Nu vom mai alege acum „Values of individual cases" ci **„Summaries for groups of cases"**, adică îi vom comunica programului să cumuleze într-un fel valorile unei variabile. Rămânem tot la graficul simplu deoarece ne interesează să reprezentăm o singură linie şi nu un grup de linii sau categorii de linii. Lansăm apoi procedura de definire a graficului prin apăsarea butonului **„Define"**.

În momentul în care am schimbat ţiunea cu care vom defini datele în zona grafică, evident, va apărea o altă fereastră de definiţie a datelor, ca în figura alăturată.

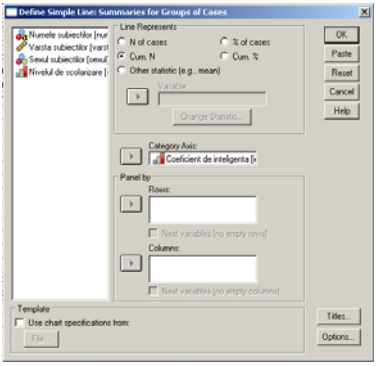


Figura 4.12 – Stabilirea parametrilor de configurare a frecvenţelor cumulate

Există multe elemente comune cu fereastra anterioară, însă şi secţiuni specifice pe care vom încerca să le descriem pe scurt. Cea mai importantă secţiune este **„Line Represents”** în care nu mai introducem pur şi simplu variabila ci definim modul în care vor fi cumulate datele. Putem opta pentru **„N of cases”** prin care SPSS va număra apariţiile fiecărei valori în şirul de date obţinând un poligon al frecvenţelor absolute pe date grupate, **„Cum. N”** adică apariţiile fiecărei valori în şirul de date (frecvenţele) dar prezentate cumulat şi exprimarea procentuală a acestora (sau frecvenţele relative şi relative cumulate) – **„% of cases”** şi **„Cum. %”**. Există şi posibilitatea să reprezentăm grafic anumite statistici folosind opţiunea **„Other statistics”** situaţie în care va trebui să includem variabila pentru care vom extrage anumiţi indicatori statistici şi de asemenea să precizăm ce indicatori se vor folosi prin activarea butonului **„Change Statistic…”**. Nu intrăm acum în detalii legate de această opţiune. În cazul nostru am ales opţiunea **„Cum. N”** deoarece dorim să reprezentăm frecvenţele cumulate absolute. Avem de stabilit şi variabila pentru care vom desena poligonul frecvenţelor cumulate absolute pe care trebuie să o

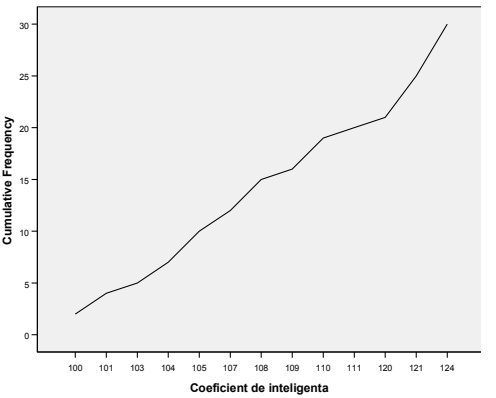


Figura 4.13 – Poligonul frecvenţelor cumulate

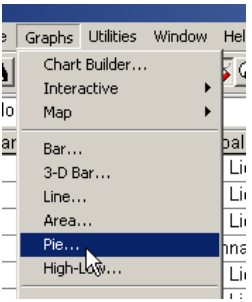


Figura 4.14 – Meniul pentru diagramele de structură

introducem în câmpul **„Category Axis:”** folosind butonul de transfer din faţa acestui câmp. Acum, nu trebuie decât să apăsăm butonul **„OK”** şi obţinem, iată, poligonul frecvenţelor absolute cumulate aşa cum ne-am dorit. Folosind celelalte trei opţiuni pot fi desenate poligoanele frecvenţelor absolute şi ale frecvenţelor relative şi relative cumulate, activităţi pe care vi le lăsăm ca exerciţiu.

Reprezentarea procentuală a unor variabile categoriale cum ar fi „genul biologic” sau „nivelul de şcolarizare” se poate face, cel mai bine, folosind un grafic de tip „plăcintă” (pie), numit şi diagramă de structură. Cu ajutorul programului SPSS vom încerca să desenăm graficul de tip plăcintă pentru variabila „nivel de şcolarizare”. Pentru a reuşi acest lucru, vom apela tot la meniul **„Graphs”** şi vom alege apoi submeniul **„Pie...”.**

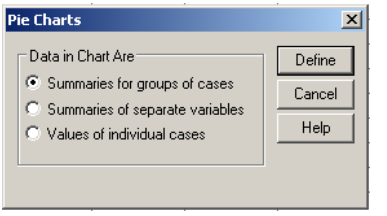


Figura 4.15 – Alegerea tipului de grafic

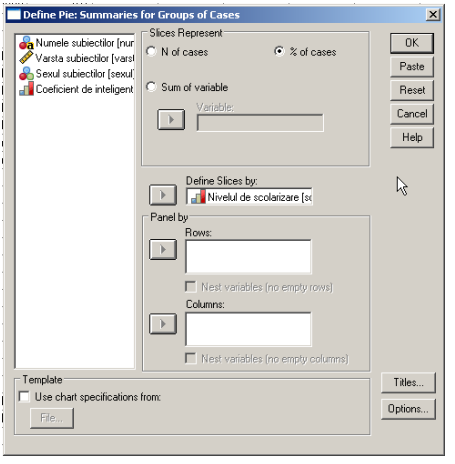


Figura 4.16 – Configurarea reprezentării prin diagramă de structură

De această dată nu mai avem ce tip de grafic plăcintă să selectăm iar următoarea fereastră prezintă doar cele trei opţiuni pe care le-am discutat. Vom selecta acum prima opţiune, **„Summaries for groups of cases”** deoarece ne interesează să reprezentăm grafic frecvenţele cu care apar subiecţii cu studii primare, cei cu gimnaziu, cu liceu, etc. şi, în plus, această reprezentare să aibă un caracter procentual. Deci, alegând această opţiu-

ne, SPSS va aduna cazurile care au o aceeaşi valoare.

Pentru a defini graficul, procedăm ca şi în exemplul anterior, apăsăm butonul **„Define”**. Se va deschide fereastra de definire a graficului, o fereastră pe care o cunoaştem deja de la reprezentarea poligonului frecvenţelor cumulate. Singura deosebire constă în secţiunea **„Slices Represents”** sau, tradus, ce reprezintă feliile plăcintei. Evident, în această secţiune am selectat opţiunea **„% of cases”** adică am comunicat programului SPSS să îmi reprezinte grafic proporţia în care se găseşte fiecare nivel al acestei variabile în eşantionul total.

Desigur, trebuie să precizăm şi variabila pentru care vor fi reprezentate grafic procentele. Acest lucru în putem face prin transferarea variabilei „Nivelul de şcolarizare” din lista variabilelor din baza de date în elementul **„Define Slices by:”** Selectaţi aşadar variabila în lista variabilelor din baza de date şi apăsaţi butonul de transfer din faţa acestui element pentru a obţine o imagine ca în figura de mai sus. În acest moment se activează şi butonul **„OK”** pe care va trebui să apăsăm pentru a lansa în execuţie procedura de desenare a graficului. După câteva momente, obţinem în fereastra de rezultate următorul grafic plăcintă:

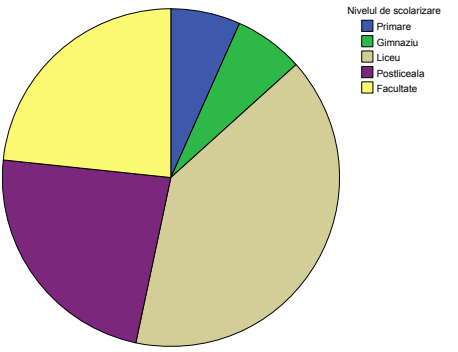


Figura 4.17 – Diagrama de structură pentru nivelul de studii

Observăm că acest grafic prezintă şi o legendă prin care suntem informaţi asupra categoriilor variabilei analizate. Putem cu uşurinţă observa că majoritatea subiecţilor sunt absolvenţi de liceu, apoi, în proporţii relativ egale absolvenţi de postliceală şi facultate. Cei mai puţini sunt cei cu studii primare şi gimnaziale.

Iată aşadar câteva metode extrem de simple prin care putem desena grafice intuitive utilizând aplicaţia SPSS. Toate aceste statistici pe care le-am prezentat şi pe care le vom mai prezenta fac parte din **analiza de date exploratorie** deoarece scopul acesteia este acela de a explora, de a cunoaşte structura datelor analizate.

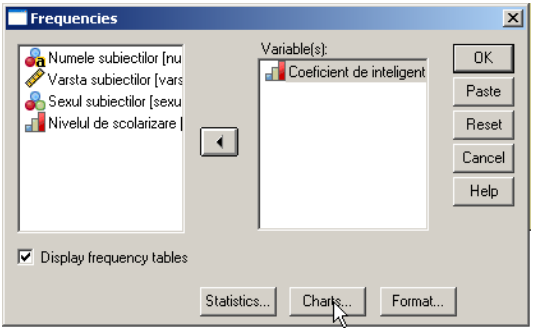


Figura 4.18 – Selectarea variabilei ce urmează să fie reprezentată

Vă mai amintiţi, desigur, fereastra studiată în capitolul anterior prin care am calculat indicatorii tendinţei centrale. Iat-o, v-o reamintesc, iar dumneavoastră va trebui să găsiţi comenzile prin care să ajungeţi la ea deoarece reprezentarea grafică a datelor se poate face şi în acest mod. Să introducem, la fel ca în capitolul anterior, variabila „Coeficient de inteligenţă" pentru a extrage indicatorii tendinţei centrale. Sper că vă amintiţi cum se face. Dacă nu, vă dau un indiciu: folosiţi butonul „Statistics…". Iată că lângă acest buton se află un alt buton, butonul **„Charts…"** pe care vom apăsa. Se deschide o nouă fereastră în care putem stabili ce grafic să desenăm pentru variabila introdusă în analiză. Opţiunea **„None"** nu desenează nici un fel de grafic şi ignoră setările grafice. Opţiunea **„Bar charts"** desenează un grafic cu bare, opţiunea **„Pie charts"** desenează grafice plăcintă şi, iată, opţiunea **„Histograms:"** ne permite desenarea unei histograme. Această opţiune o întâlnim aici, alături de caseta de bifare **„With normal**

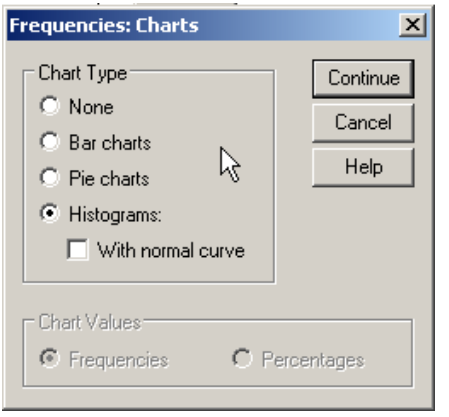


Figura 4.19 – Modul de reprezentare grafică din analiza frecvenţelor

curve” ce permite desenarea curbei normale care caracterizează datele din variabila noastră. Nu vom bifa acum această casetă deoarece despre curba normală vom discuta într-un alt capitol. Ne rezumăm să alegem doar graficul sub formă de histogramă.

Anumite grafice pot fi reprezentate sub formă de frecvențe absolute (opțiunea **„Frequencies”**) sau sub formă de frecvențe relative (opțiunea **„Percentages”**) prin selectarea uneia dintre cele două opțiuni din secțiunea **„Chart Values”.** Folosind această fereastră avem de configurat mult mai puține elemente. Deoarece am introdus anterior variabila pe care dorim să o analizăm, SPSS va lua singur decizia modului în care va desena graficul. Apăsăm butonul **„Continue”** pentru a confirma acțiunea noastră şi apoi butonul **„OK”** pentru a lansa analiza.

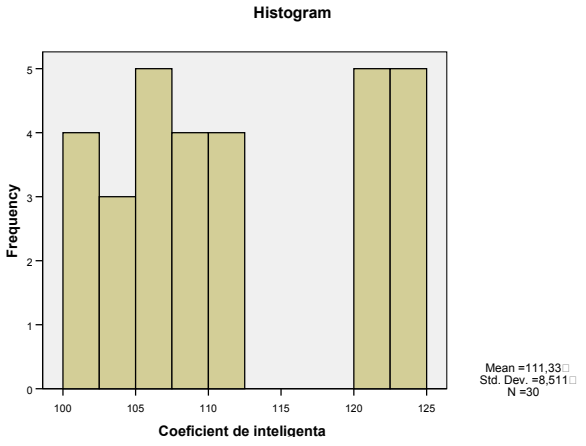


Figura 4.20 – Histograma pentru coeficientul de inteligență

După câteva momente, SPSS va extrage indicatorii tendinței centrale sub forma celor două tabele discutate în capitolul anterior, însă ne va oferi şi histograma variabilei „Coeficient de inteligență” alături de o serie de date din inventarul statistic de bază pe care le vom discuta într-un alt capitol.

IV.2 Analiza grafică a tendinței centrale

Afirmam în capitolul dedicat indicatorilor tendinței centrale că media este cel mai „fin”, cel mai precis indicator însă ea are o problemă – este sensibilă la scoruri extreme şi în acest caz nu mai aproximează bine tendința

centrală a setului de date. Totuşi, cum ne dăm seama de existenţa acestor scoruri extreme? Tot analiza exploratorie ne vine în ajutor. Utilizând reprezentările de date de mai sus, ne putem face o imagine intuitivă asupra acestor scoruri extreme. Există oare o metodă mai precisă de a le depista? Răspunsul este pozitiv şi vom prezenta în continuare două tipuri de grafice deosebite, dar care pot caracteriza mai precis scorurile extreme.

IV.2.1 Graficul „tulpină şi frunze”

A fost conceput de către Tuckey (1977) şi se aseamănă foarte mult cu histogramele de frecvenţe, însă, spre deosebire de acestea, afişează şi scorurile individuale. Un asemenea grafic se desenează mai uşor manual decât computerizat şi oferă o imagine precisă a distribuţiei deoarece alături de reprezentarea numerică oferă şi o imagine vizuală a datelor analizate (Dancey, şi alţii, 2002).

Dacă avem, de exemplu, următorul şir ordonat de date: 2, 12, 12, 19, 19, 20, 20, 20, 25 şi dorim să construim un grafic „tulpină şi frunze”, acesta va arăta după cum urmează:

Tulpina	*Frunzele*
0	*2*
1	*2299*
2	*0005*

Oare ce informaţii noi ne aduce acest grafic şi cum se interpretează? Destul de simplu. „Tulpina” reprezintă valorile „zecilor” iar „frunzele” valorile „unităţilor”. Deci, tulpina 0 cuprinde scorurile de la 0 la 9, tulpina 1 scorurile de la 10 la 19, tulpina 2 scorurile de la 20 la 29 şi aşa mai departe. Frunzele reprezintă frecvenţele scorului observat. Astfel la tulpina 0 (scoruri între 0 şi 9) avem o singură apariţie a lui 2. La tulpina 1 (scoruri între 10 şi 19) avem frunza 2 de două ori (adică două apariţii ale lui 12) şi frunza 9 de două ori (două apariţii ale lui 19). La tulpina 2 (scoruri între 20 şi 29) avem frunza 0 cu trei apariţii (trei scoruri de 20) şi frunza 5 cu o singură apariţie

(un singur scor de 25). Simplu, nu? Iată că, deşi seamănă cu o histogramă, acest grafic adaugă în plus posibilitatea monitorizării scorurilor individuale. Pentru a înțelege mai bine, să considerăm un set mai voluminos de date: *1, 1, 2 ,2 ,2 , 5, 5, 5, 12, 12, 12, 12, 14, 14, 14, 14, 15, 15, 15, 18, 18, 24, 24, 24, 24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 28, 28, 28, 28, 28, 28, 28, 32, 32, 33, 33, 33, 33, 34, 34, 34, 34, 34, 35, 35, 35, 35, 35, 42, 42, 42, 43, 43, 44*. Pentru acest set de date avem următorul grafic „tulpină şi frunze"

0	*11222555*
1	*2222444455588*
2	*44444455555558888888*
3	*22333344444455555*
4	*222334*

Iată că prin această reprezentare avem o metodă mai concisă şi mai sintetică de explorare a unui volum mare de date. Putem observa că scorurile de la tulpina 2 (între 20 şi 29) sunt cele mai frecvente iar datele sunt relativ grupate, fără scoruri extreme. În acest caz am putea decide utilizarea mediei ca indicator al tendinței centrale. Comparativ, să considerăm un şir de date cu scoruri extreme şi să îl reprezentăm sub forma acestui grafic.

0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 13, 13, 13, 13, 13, 18, 18, 18, 23, 64

Graficul pentru acest şir de date ar fi:

0	*000002222222233333333355555555555555777777777777799999999*
1	*000000033333888*
2	*3*
6	*4*

Observăm că graficul ne indică faptul că majoritatea scorurilor se situează sub 20 însă avem şi un scor extrem, scorul 64. Iată deci o situație tipică în care mediana va fi cel mai bun indicator al tendinței centrale. Totuşi, acest grafic efectuat pe blocuri de 10 valori nu este foarte informativ. Nu avem o reprezentare foarte clară a scorurilor sub 20 deoarece există foarte

multe. Prin urmare ar fi mai comod pentru noi să reprezentăm grafic aceste scoruri folosind nu blocuri de 10 ci blocuri de cinci (între 0 şi 4, apoi între 5 şi 9, între 10 şi 14 şi aşa mai departe). Folosind un asemenea sistem, graficul devine:

0. *00000222222223333333333*
*0** *55555555555555577777777777799999999*
1. *000000033333*
*1** *888*
2. *3*
6. *4*

Observăm că am folosit punctul (.) pentru a desemna prima jumătate a blocului (0-4) şi steluţa (*) pentru a desemna a doua jumătate a blocului (5-9). De data aceasta lucrurile au devenit ceva mai clare. Există acelaşi scor extrem, 64, însă acum putem afirma precis că cei mai mulţi subiecţi au obţinut rezultate între 5 şi 9 ceea ce reprezintă un plus de informaţie. Rolul acestei reprezentări este mai ales acela de a identifica existenţa scorurilor extreme într-un set de date în vederea stabilirii indicatorului tendinţei centrale ce va fi folosit.

Discutam, pe la începutul lucrării noastre, despre situaţia în care ordonăm toate valorile posibile şi apoi stabilim frecvenţele absolute prin numărarea apariţiilor fiecărui scor individual. La fel, am abordat şi problema organizării datelor în clase. Am văzut atunci că un număr mare de clase reprezintă o abordare analitică, accentul căzând mai mult pe valori individuale şi se pierde tendinţa de ansamblu. Analog, un număr mic de clase determină o perspectivă sintetică, unde vedem clar tendinţa de ansamblu, însă de pierde viziunea individuală. Unul dintre avantajele incontestabile ale acestui grafic este acela că permite o viziune de ansamblu fără însă a se pierde repartiţia scorurilor individuale în distribuţie.

Prin intermediul acestui tip de grafic se pot reprezenta, comparativ, două distribuţii, în situaţia în care dorim, spre exemplu, să comparăm rezulta-

tele la două probe psihologice. Utilizând cele două seturi de date de mai sus, am putea să le reprezentăm grafic după cum urmează:

Test 1		Test 2
99999999777777777777555555555555555333333332222222200000	**0**	11222555
888333330000000	**1**	2222444455588
3	**2**	44444455555558888888
	3	22333344444455555
	4	222334
	5	
4	**6**	

Datorită uşurinţei în realizare şi avantajelor pe care le oferă, acest tip de grafic se utilizează tot mai frecvent în ştiinţele socio-umane.

IV.2.2 Graficul „cutie" (box-plot)

Deşi graficul anterior sesizează scorurile extreme suficient de bine, uneori devine destul de puţin sensibil iar scorurile extreme sunt greu de observat. O metodă mai precisă o reprezintă graficul „cutie" sau graficul „cutiei cu mustăţi". Această tehnică, la fel ca şi tehnica anterioară a fost descoperită de către Turkey (1977) şi reprezintă unele dintre cele mai importante metode de analiză exploratorie a datelor. La fel ca şi graficul „tulpină şi frunze", graficul „cutie" descrie modul în care sunt distribuite datele şi oferă, în plus, o metodă precisă de depistare a scorurilor extreme (Dancey, şi alţii, 2002).

Acest tip de reprezentare este formată dintr-un dreptunghi cu baza inferioară situată în dreptul percentilei 25 iar baza superioară în dreptul percentilei 75, cuprinzând între aceste limite, 50% dintre rezultate (în fapt, intervalul cuartil despre care vom vorbi în capitolul următor). În interiorul acestui dreptunghi se află marcată linia mediană, corespunzătoare percentilei 50. Distanţa dintre cele două baze (intervalul cuartil) poartă numele de *h-spread* sau pe scurt H. De la cele două baze pornesc, în sus şi în jos, două „mustăţi" de lungime maximă 1,5 H. Între limitele acestor două mustăţi se află zona normală. Orice scor mai mare de limita „mustăţii" superioare sau mai mic de limita „mustăţii" inferioare se consideră a fi un scor extrem.

Pentru realizarea acestui grafic vom parcurge următorii paşi considerând şirul ordonat de date 2, 12, 12, 19, 19, 20, 20, 20, 25

Pasul 1 – Stabilirea medianei şi a poziţiei pe care aceasta o ocupă în cadrul şirului de date. În cazul nostru mediana este 19 iar poziţia pe care o ocupă este poziţia a cincia.

2, 12, 12, 19, 19, 20, 20, 20, 25

Pasul 2 – Calculul „pivoţilor". Pivoţii reprezintă valorile care împart şirul de date în sfertul inferior (primele 25% dintre scoruri, numit şi **cuartilul inferior**) şi sfertul superior (ultimele 25% dintre scoruri, numit şi **cuartilul superior**). Dacă mediana este valoarea care împarte şirul ordonat de date în două părţi egale, cuartilul este valoarea ce împarte acelaşi şir în 4 părţi egale, după cum vom vedea în capitolul ce urmează. Reconsiderând exemplul, obţinem poziţia primului cuartil între a doua şi a treia valoare, adică primul cuartil va avea valoarea 12 (amintiţi-vă exemplul de calcul al medianei). Similar, al treilea cuartil va fi situat între poziţia 7 şi poziţia 8 şi va avea valoarea 20.

2, 12, 12, 19, 19, 20, 20, 20, 25

Pivoţii. În cazul exemplului de faţă, vor avea valorile 12 şi 20. Putem acum trasa dreptunghiul care ne interesează (box).

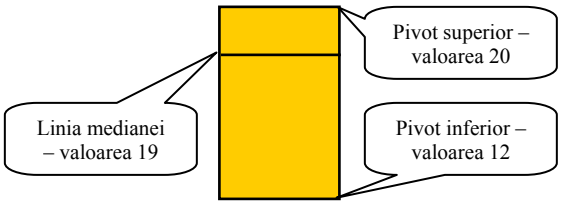


Figura 4.21 – Calculul pivoţilor

Pasul 3 – Obţinerea valorii expansiunii (factorul *h-spread*). Acest factor nu reprezintă decât amplitudinea scorurilor dintre cei doi pivoţi (intervalul cuartil). În cazul nostru, *h-spread = 20 – 12 = 8*

Pasul 4 – Stabilirea scorurilor extreme. Un scor extrem este un scor care depăşeşte o dată şi jumătate

valoarea factorului *h-spread* raportat la cei doi pivoți. Aceste puncte de referință se numesc *limite de barieră*. În cazul nostru, 1,5 x 8 = 12. Deci limitele de barieră sunt 12 – 12 = 0 şi 20 + 12 = 32. Scorurile mai mici de 0 vor fi considerate extreme deoarece depăşesc limita de barieră inferioară în jos iar scorurile mai mari de 32 vor fi considerate de asemenea extreme deoarece depăşesc în sus limita de barieră superioară.

Pasul 5 – Scorurile care se află între pivoți şi limitele de barieră, mai apropiate de limitele de barieră se numesc ***scoruri adiacente***. În cazul nostru scorurile adiacente sunt 2 şi 25 deoarece 2 se află între 0 şi 12 iar 25 este situat între 20 şi 32.

2, 12, 12, 19, 19, 20, 20, 20, 25

Liniile cuprinse între pivoți şi bariere (zonele scorurilor adiacente) poartă numele de „mustăți" iar dreptunghiul format de cei doi pivoți care conține mediana se numeşte „cutie". De aici provine şi numele graficului, grafic „cutie" sau „cutie cu mustăți". Zona „mustăților" reprezintă de fapt amplitudinea scorurilor care părăsesc zona aşa numită normală delimitată de cei doi pivoți, însă rămân în limitele barierelor. Orice scor care iese din limita barierelor se numeşte scor extrem.

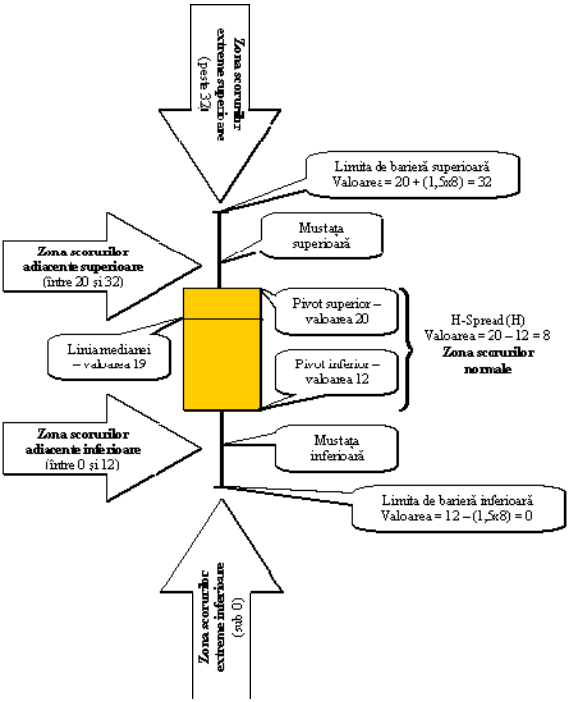

Zona scorurilor extreme superioare (peste 32)
Limita de barieră superioară
Valoarea = 20 + (1,5x8) = 32
Mustața superioară
Zona scorurilor adiacente superioare (între 20 și 32)
Pivot superior – valoarea 20
H-Spread (H)
Valoarea = 20 – 12 = 8
Zona scorurilor normale
Linia mediană – valoarea 19
Pivot inferior – valoarea 12
Zona scorurilor adiacente inferioare (între 0 și 12)
Mustața inferioară
Limita de barieră inferioară
Valoarea = 12 – (1,5x8) = 0
Zona scorurilor extreme inferioare (sub 0)

Figura 4.22 – Graficul box-plot

Dacă reconsiderăm şirul anterior de date, 0, 0, 0, 0, 0, 2,

2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 13, 13, 13, 13, 13, 18, 18, 18, 23, 64, atunci graficul nostru se va prezenta ca în figura de mai jos.

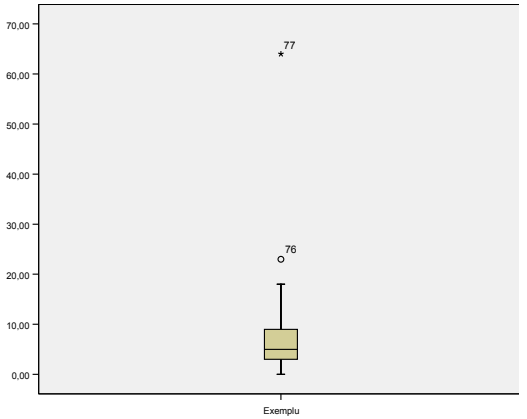


Figura 4.23 – Graficul box-plot în SPSS. Se observă scorul extrem, al 77-lea element

Iată că observăm existenţa a două scoruri extreme însă…. acestea sunt notate cu 77 şi 76. Am fi înclinaţi să credem că aceste valori reprezintă scorurile înseşi. Totuşi să nu uităm că acest grafic se bazează pe poziţia unui scor în şirul ordonat de date. Prin urmare, 77 şi 76 nu reprezintă scorurile propriu zise ci poziţia acestora în şirul ordonat de date. Deci poziţia 77 se referă la ultimul scor, adică la valoarea 64 iar poziţia 76 la penultimul scor, adică la valoarea 23. Avem aşadar două scoruri extreme în şirul nostru de date şi anume scorul 23 şi 64. Desigur, în acest caz indicatorul ce exprimă cel mai bine tendinţa centrală este mediana. Programul SPSS oferă şi o modalitate de a marca scorurile extreme. Observăm că alături de poziţia 77 (corespunzătoare scorului 64) apare o steluţă. Această steluţă are rol de avertisment, în sensul că valoarea 64 este foarte îndepărtată de limita superioară a şirului de date. De asemenea, alături de poziţia 76 (corespunzătoare scorului 23) apare un cerculeţ. Acest cerculeţ de spune că, deşi 23 este şi el un scor extrem, nu este atât de depărtat de limita superioară a şirului de date.

IV.3 Tratarea scorurilor extreme

Nu ne putem mulțumi să constatăm doar existența acestor scoruri extreme, mai ales în condițiile în care acestea ar afecta puternic analiza datelor. Un scor extrem duce la o distribuție de date care nu mai respectă criteriul normalității, face media un indicator irelevant pentru acel set de date şi, implicit, determină imposibilitatea utilizării statisticilor parametrice. Totuşi, de unde pot să apară aceste scoruri extreme? Ce surse există care pot determina prezența unor asemenea valori? (Popa, 2003)

- Cea mai frecventă sursă se referă la **erorile de înregistrare** a datelor. Dacă introducem un chestionar, la care subiecții răspund la întrebări pe o scală de la 1 la 5 şi în loc să tastăm 5, tastăm, din greşeală, 55, evident că acesta va apărea ca fiind un scor extrem. Din fericire, remedierea acestei probleme este simplă. Nu trebuie decât să analizăm minimul şi maximul distribuției. Dacă una dintre aceste valori iese din domeniul de definiție, putem reveni uşor la baza de date şi o vom corecta. Dacă sunt mai multe valori incorecte, putem sorta datele crescător sau descrescător şi putem corecta mai uşor erorile sau putem apela la o operațiune de recodificare;
- **Erorile de eşantion** se referă la situația în care eşantionul estras este prea mic iar posibilitatea de apariție a unui caz atipic (extrem) va deveni foarte mare dar şi la cazul în care eşantionul a fost extras dintr-o populație asimetrică. Putem, de exemplu, să studiem înălțimea unor subiecți dintr-un liceu iar pentru aceasta extragem zece persoane din două clase. Probabilitatea ca un elev să fie baschetbalist este destul de mare. Dacă avem „ghinionul" să îl selectăm tocmai pe acela, va apărea, evident, şi un caz extrem. De asemenea, dacă una dintre clase este reprezentată de clasa de sport, cu elevi înalți, printre care am gă-

sit şi un elev cu înălţime normală, acea înălţime normală se va comporta ca un caz extrem;

- **Erorile determinate de existenţa unor populaţii diferite de valori** apar atunci când în eşantionul studiat găsim, sub aspectul variabilei măsurate, mai multe grupe de subiecţi care diferă între ele. De exemplu, dacă studiem veniturile angajaţilor unei întreprinderi, avem suficiente şanse ca salariile conducerii să se comporte ca scoruri extreme în raport cu celelalte categorii de angajaţi. În realitate, acestea nu sunt scoruri extreme ci problema se referă tocmai la faptul că există mai multe categorii de salarizare în acea instituţie;
- **Erorile de măsurare** se datorează unor măsurători incorecte sau de apariţia unor situaţii anormale în condiţiile experimentale. La înregistrarea temperaturii unor persoane, se poate defecta, la un moment dat, aparatul de înregistrare sau, măsurând performanţele unor sportivi, putem constata apariţia unor valori extreme ca efect al dopajului;

Indiferent de sursa de eroare, înaintea realizării prelucrărilor statistice prin metode şi tehnici specifice, este absolut necesară analiza preliminară a datelor, prin intermediul căreia să depistăm caracteristicile distribuţiei iar dacă acestea sunt inadecvate prelucrărilor necesare, să impunem măsuri corective. Din fericire, cu excepţia situaţiei defectării unor aparate sau a designurilor de cercetare defectuoase, există suficiente mijloace prin care putem „repara" aceste disfuncţionalităţi. Deoarece tehnicile de transformare a datelor brute presupun şi alte cunoştinţe, vom amâna tratarea acestora pentru un capitol viitor.

Orice prelucrare mai avansată de date începe cu statisticile descriptive şi este absolut necesară o asemenea analiză exploratorie pentru a vedea care sunt caracteristicile şirului nostru de date şi ce indicatori putem folosi. Pro-

gramul SPSS ne pune la dispoziție, iată, o serie de metode prin intermediul cărora putem verifica rapid şi precis structura datelor cu care lucrăm.

IV.3.1 Obținerea graficelor pentru tendința centrală în SPSS

Vom rămâne la baza noastră de date IQ şi ne propunem să obținem aceste grafice folosind programul SPSS. Pentru a realiza acest lucru, apelăm la meniul **„Analyze”**, apoi la submeniul **„Descriptive Statistics”** şi în final utilizăm opțiunea **„Explore…”** Iată o nouă ancoră mnemotehnică care ne spune că ne aflăm în analiza datelor, la statistici descriptive şi la analize exploratorii. Efectuând clic pe această opțiune, se deschide o fereastră nouă prin intermediul căreia putem stabili analizele de date ce vor fi efectuate.

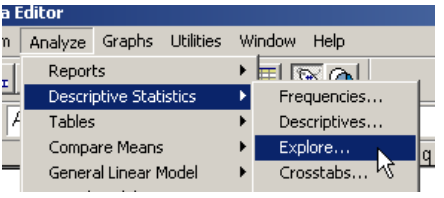


Figura 4.24 – Meniul de analiză exploratorie a datelor

Avem mai multe elemente noi în această fereastră. În primul rând, caseta **„Dependent List:”** care se referă la variabilele ce vor fi incluse spre analiză. În această casetă vom include variabila noastră şi anume „Vârsta subiecților” prin selectarea ei din lista variabilelor din baza de date şi apăsarea butonului de transfer. În acest moment se activează şi butonul **„OK”** ceea ce ne spune că putem face imediat prelucrarea datelor. Să nu ne grăbim însă şi să studiem mai atent această fereastră. Într-un capitol anterior discutam despre faptul că indicatorii tendinței centrale (şi nu numai ei) pot fi obținuți prin mai multe metode. Atunci am analizat numai opțiunea „Fre-

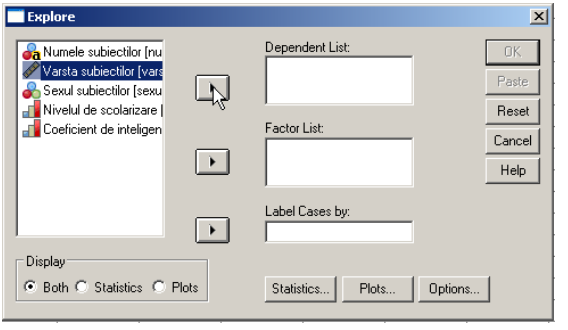


Figura 4.25 – Configurarea variabilelor în analiza descriptivă exploratorie

quencies…" şi am promis că celelalte modalităţi vor fi studiate cu altă ocazie. Iată că a venit şi momentul acesta. Caseta **„Factor List:"** se referă la le categoriale ce pot fi utilizate în vederea grupării analizei în funcţie de anumite criterii. Atât această listă cât şi elementul **„Label Cases by:"** vor fi studiate atunci când vom clarifica conceptele de variabilă dependentă şi abilă independentă. În secţiunea **„Display"** situată imediat sub lista variabilelor din baza de date putem stabili ce fel de rezultate să se afişeze în fereastra de rezultate. Putem opta pentru afişarea doar a rezultatelor prelucrărilor statistice (**„Statistics"**), pentru afişarea doar a graficelor (**„Plots"**) sau a ambelor (**„Both"**). Noi vom dori să afişăm toate aceste date şi de aceea am selectat opţiunea **„Both".** Butonul **„Statistics…"** permite afişarea statisticilor descriptive ce vor fi calculate. La apăsarea acestuia va apărea următoarea fereastră:

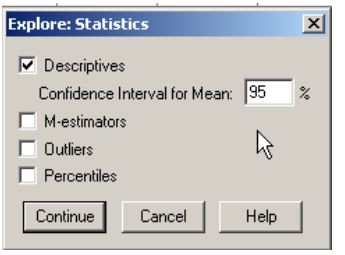


Figura 4.26 – Stabilirea indicatorilor statistici

Sunt mai multe opţiuni aici, noi vom bifa doar opţiunea de calcul a statisticilor descriptive, **„Descriptives"**. Secţiunea **„Confidence interval for Mean:"** se referă la intervalul de încredere la care va fi calculată media. Asupra acestui aspect vom reveni însă într-un alt capitol. De asemenea, celelalte opţiuni nu ne interesează deocamdată, astfel încât le vom ignora pe moment. Menţionăm doar că **„M-estimators"** reprezintă o alternativă la media şi mediana eşantionului, **„Outliers"** afişează primele cinci valori cele mai mari, respectiv cele mai mici (extremele) iar **„Percentiles"** permite calculul unor valori percentile. Toate aceste lucruri însă vor fi studiate cu altă ocazie. În vederea confirmă-

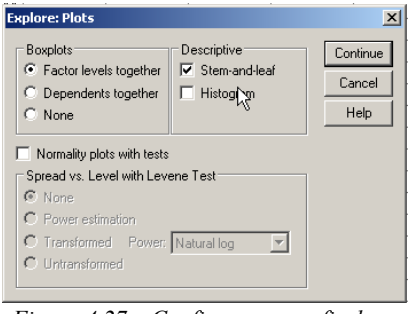


Figura 4.27 – Configurarea graficelor exploratorii

rii acțiunii noastre, va trebui să apăsăm butonul **„Continue”** pentru a reveni la fereastra anterioară.

Acționarea butonului **„Plots…”** determină afişarea unei alte ferestre ce va configura modul de prezentare grafică a datelor. Şi aici avem mai multe informații. Pentru început observăm că este bifată caseta **„Stem-and-leaf”** adică va fi afişat graficul „tulpină şi frunze”.

De asemenea vom afişa şi histograma, bifând caseta **„Histogram”** Secțiunea **„Boxplots”**, situată în partea stângă-sus permite configurarea modului de calcul al graficului „cutie cu mustăți” discutat anterior. Putem opta pentru combinarea nivelurilor variabilei categoriale (dacă am inclus vreuna în fereastra anterioară) alegând **„Factor levels together”** sau putem combina mai multe variabile incluse în lista variabilelor independente alegând opțiunea **„Dependents together”**. Desigur, putem renunța la afişarea acestui grafic selectând opțiunea **„None”.** În cazul nostru, având doar o singură variabilă inclusă spre analiză, oricare dintre primele două opțiuni le-am alege obținem acelaşi efect. Părăsirea acestei casete va fi făcută tot prin apăsarea butonului **„Continue”**. Nu ne rămâne acum decât să apăsăm butonul **„OK”** pentru a lansa procedurile de analiză. În fereastra de rezultate avem acum mai multe tabele şi grafice.

Tabel 4.1 – Sumarul cazurilor

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Varsta subiectilor	30	100,0%	0	,0%	30	100,0%

În tabelul **„Case Processing Summary”** ne sunt prezentate numărul de cazuri. Aflăm că avem 30 de cazuri valide, adică 30 de subiecți ce au date completate pentru variabila „Vârsta subiecților” ceea ce reprezintă un procent de 100%. Evident, nu avem nici un caz lipsă (Missing) şi, corespunzător, un procent de 0%. În total sunt deci 30 de cazuri ce reprezintă un procent de 100%.

Următorul tabel este acela al statisticilor descriptive (**„Descriptives”**). Remarcăm cu uşurinţă media (**Mean**) egală cu 29,33 ani şi mediana (**Median**) egală cu 29 de ani. Sunt mult mai multe informaţii aici, legate de eroarea standard a mediei, intervalul de încredere al acesteia precum şi indicatori ai dispersiei pe care-i vom studia în alt capitol. Observăm că din acest tabel lipseşte modul. Într-adevăr, folosind această opţiune nu putem calcula modul.

Tabel 4.2 – Statistici descriptive

Descriptives

			Statistic	Std. Error
Varsta subiectilor	**Mean**		**29,33**	**1,008**
	95% Confidence Interval for Mean	Lower Bound	27,27	
		Upper Bound	31,40	
	5% Trimmed Mean		29,33	
	Median		**29,00**	
	Variance		30,506	
	Std. Deviation		5,523	
	Minimum		20	
	Maximum		39	
	Range		19	
	Interquartile Range		10	
	Skewness		,116	,427
	Kurtosis		-,878	,833

Urmează apoi histograma pentru variabila „Vârsta subiecţilor” însoţită de o serie de date statistice cum ar fi media, abaterea standard şi numărul de cazuri şi, iată, graficul „tulpină şi frunze” pe care l-am discutat anterior.

```
Varsta subiectilor Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

     3,00        2 .  001
      ,00        2 .
     5,00        2 .  44444
     3,00        2 .  666
     6,00        2 .  899999
     4,00        3 .  0111
     1,00        3 .  2
     3,00        3 .  455
     2,00        3 .  77
     3,00        3 .  889

Stem width:    10
Each leaf:     1 case(s)
```

Observăm că acest grafic a fost creat din doi în doi (20-21, 22-23, 24-25, 26-27, 28-29 ani şi aşa mai departe) pentru a oferi o bună reprezentare a datelor. Iată că, analizând acest grafic, putem afirma că cei mai mulţi subiecţi

au vârste de 28 şi 29 de ani şi nu există scoruri extreme. Pentru a fi siguri de acest lucru, privim următorul grafic, graficul „cutie".

Privind şi acest grafic constatăm lipsa scorurilor extreme, mediana fiind situată la valoarea 29 şi o distribuție destul de uniformă a acestor rezultate.

Putem afirma, deocamdată, că aceste date permit utilizarea mediei ca indicator al tendinței centrale pentru setul nostru.

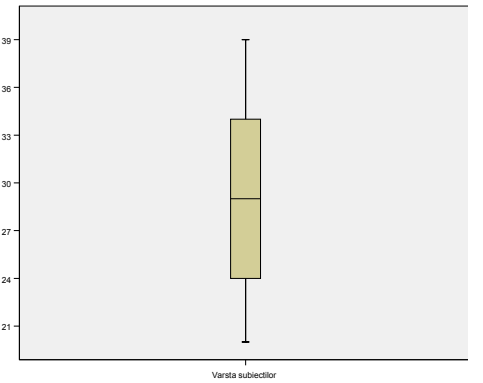


Figura 4.28 – Graficul box-plot

Concluzii:

- *Media unei populații la un parametru poate fi aproximată prin media eşantioanelor extrase din populație;*
- *Analiza scorurilor extreme se poate face în baza a două tipuri principale de grafice: graficul „tulpină şi frunze" şi graficul tip „cutie"*
- *Prezența scorurilor extreme scade puterea statistică a analizei. Prin urmare trebuie să efectuăm analize statistice preliminare asupra setului de date şi eventual să procedăm la tehnici de transformare a datelor înaintea analizei statistice efective.*

V. STATISTICI DESCRIPTIVE – ÎMPRĂŞTIEREA

În acest capitol se va discuta despre:

- Conceptul de împrăştiere;
- Indicatori simpli şi complecşi ai abaterii de la tendinţa centrală;
- Calculul indicatorilor abaterii de la tendinţa centrală;

După parcurgerea capitolului, cursanţii vor fi capabili să:

- Înţeleagă legătura dintre tendinţa centrală şi dispersie;
- Calculeze indicatorii abaterii de la tendinţa centrală;
- Utilizeze SPSS in calculul indicatorilor abaterii de la tendinţa centrală.

Indicatorii tendinţei centrale caracterizează, după cum am văzut, valoarea în jurul căreia se grupează datele. Ei au, desigur, o mare putere informativă, însă nu ne prezintă o măsură a compoziţiei acestor date. De exemplu, media greutăţii unui număr de 2 copii este 20 Kg. Cei doi copii pot să aibă 20 de Kg fiecare, unul 10 Kg şi celălalt 30 Kg, sau unul 15 Kg iar celălalt 25 Kg şi aşa mai departe. Iată deci că indicatorii tendinţei centrale nu pot caracteriza singuri o distribuţie. Avem nevoie, pentru aceasta, de o măsură a gradului de răspândire, de împrăştiere a datelor în jurul tendinţei centrale, cum se repartizează rezultatele în jurul acestei valori. Aceşti indicatori, care măsoară gradul de împrăştiere a rezultatelor în jurul tendinţei centrale, poartă numele de **indicatori ai împrăştierii**. Indicatorii împrăştierii pot fi grupaţi în indicatori elementari ai împrăştierii şi indicatori sintetici ai împrăştierii.

Modelul statistic al mediei va trebui aşadar completat cu alţi indicatori caracteristici gradului de împrăştiere a rezultatelor în jurul tendinţei centrale. Dacă reluăm exemplul cu prietenii mei şi numărul de fraţi pe care fiecare dintre ei îi au, am stabilit că, primul prieten are un frate, al doilea are doi fraţi, al treilea şi al patrulea au câte trei fraţi şi al cincilea are patru fraţi. Re-

zultase o medie de 2,6 fraţi, cu alte cuvinte doi fraţi întregi şi unul fără un picior. Fiind un model statistic, media estimează o realitate. Să vedem acum dacă această medie estimează corect sau nu realitatea fraţilor prietenilor mei.

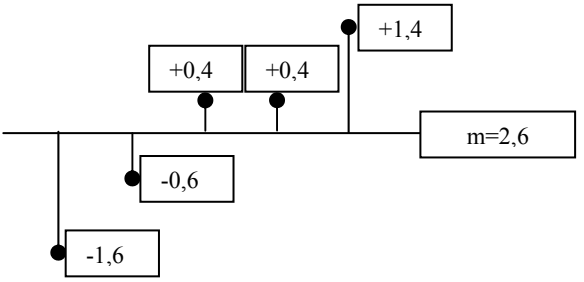


Figura 5.1 – Diferenţele dintre scorul observat şi medie

În figura 5.1 am reprezentat grafic distanţele la care se situează fiecare scor particular faţă de medie. Este aşadar media un bun model pentru numărul de fraţi ai primului prieten? Primul meu prieten are un singur frate, iar media postulează 2,6 fraţi. Iată că, în ceea ce-l priveşte pe primul prieten, media **supraestimează** modelul real cu 1,6 fraţi. Pentru al doilea prieten, media supraestimează realitatea cu doar 0,6 fraţi (probabil o mână şi un picior dintr-un frate). Referitor la ceilalţi trei prieteni, media **subestimează** realitatea cu 0,4 fraţi şi 1,4 fraţi.

Mărimea acestor supraestimări, respectiv subestimări se obţine foarte simplu, scăzând din medie scorul particular (x-m). Cum am putea folosi aceste informaţii pentru a testa acurateţea modelului? O variantă ar fi să adunăm toate aceste abateri pentru a vedea suma, totalul abaterilor elementelor de la modelul statistic al mediei.

$$(-1,6)+(-0,6)+(0,4)+(0,4)+(1,4)=0$$

Am obţinut totalul erorilor, totalul abaterilor de la medie ca fiind zero. Cu alte cuvinte, modelul nostru statistic reprezintă perfect realitatea? Aşa să fie? Ei bine, nu, deoarece tocmai am demonstrat particularitatea esenţială a mediei, conform căreia scorurile mai mici decât media balansează perfect

scorurile mai mari decât media. Totuşi, cum putem obține mărimea erorii? Cum putem estima dacă modelul statistic este o imagine suficient de bună pentru a reprezenta realitatea? Vom vedea imediat. Pentru început să ne concentrăm asupra unor indicatori elementari ai împrăştierii.

V.1 Indicatori elementari ai împrăştierii

Indicatorii elementari ai împrăştierii țin cont doar de câteva valori iar calculele sunt efectuate în baza acestora. Deşi sunt foarte uşor de obținut, sunt sensibili modul de distribuție a rezultatelor în colecția de date.

V.1.1 Amplitudinea de variație

Termenul englezesc pentru acest indicator este acela de „Range" iar noi l-am abordat deja în momentul în care am făcut sistematizarea datelor, doar că nu ştiați încă (deşi probabil bănuiați) că reprezintă un indicator al împrăştierii. Amplitudinea de variație nu este altceva decât diferența dintre valoarea maximă şi valoarea minimă din şirul nostru de date. Prin urmare, $A = X_{max} - X_{min}$. Dacă, avem, de exemplu, următoarele date: 10, 22, 31, 9, 24, 27, 29, 9, 23, 12 atunci constatăm cu uşurință că valoarea cea mai mare din şir este 29 iar cea mai mică valoare este 9. În acest caz amplitudinea de variație devine $A = 29 - 9 = 20$.

Ştim, de asemenea, semnificația acestei valori. Între maxim şi minim pot exista cel mult 20 de valori diferite. Amplitudinea de variație indică, în mod absolut, domeniul de valori între care este cuprinsă distribuția. Amplitudinea de variație poate fi exprimată şi procentual, ca expresie a raportului dintre amplitudinea de variație şi medie.

$$\boldsymbol{A\% = \frac{A}{\bar{x}} \times 100} \qquad \text{(formula 5.1)}$$

În cazul nostru, $\boldsymbol{A\% = \frac{20}{19{,}6} \times 100 = 102\%}$. Acest indicator poartă numele de **amplitudine relativă de variație.** În exemplul nostru, avem 10

valori iar media acestora este 19,6. Amplitudinea relativă de variație este, în cazul nostru, 102%. Amplitudinea relativă de variație o putem utiliza în condițiile în care cunoaştem domeniul teoretic între limitele căruia se încadrează distribuția. Putem astfel să comparăm amplitudinea reală de variație a datelor cu amplitudinea de variație teoretică.

Indicatorii amplitudinii de variație pot fi utilizați atunci când cunoaştem plaja de variație normală a unui fenomen, deoarece apariția unor valori extreme duce la amplitudini de variație aberante. În exemplul nostru, dacă introducem o singură valoare în plus, 120, atunci minimul rămâne 9 iar maximul devine 120. Amplitudinea de variație va fi 120 – 9 = 111, desigur o valoare suspectă pentru şirul nostru de date.

Analizând figura alăturată, observăm că cele două distribuții au aceeaşi amplitudine de variație. Totuşi, forma lor diferă foarte mult. În timp ce distribuția „A" este puternic eterogenă, distribuția „B" este mult mai omogenă.

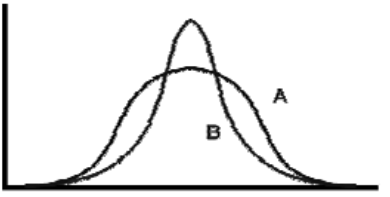


Figura 5.2 – Două distribuții cu aceeaşi amplitudine

V.1.2 Abaterea cuartilă sau abaterea intercuartilă

Este un alt indicator elementar al împrăştierii, independent de existența unor valori extreme, fiind, prin urmare, mai stabil şi mai precis în comparație cu amplitudinea de variație. Ce este însă un cuartil? Am mai adus discuția despre acest subiect în capitolul anterior. Dacă ştiți ce este mediana şi dacă vă spunem că mediana nu este altceva decât al doilea cuartil, atunci cred ca puteți deduce semnificația termenului de cuartil. Ştim că mediana este valoarea care împarte setul ordonat de date în două părți egale; jumătate dintre valori sunt mai mici decât mediana, jumătate dintre valori sunt mai mari decât mediana. Ei bine, **valorile cuartile sunt valorile care împart acelaşi set ordonat de date în patru părți egale**. Un sfert din valori (25%) sunt mai mici decât primul cuartil (Q_1), două sferturi din valori (jumătate) sunt mai

mici decât al doilea cuartil (Q_2 – iată că, de fapt, Q_2 nu este altceva decât mediana) iar trei sferturi din valori (75%) sunt mai mici decât al treilea cuartil (Q_3). **Abaterea cuartilă nu este altceva decât diferența dintre cuartilul 3 şi cuartilul 1**.

Prin urmare, $R_Q=Q_3 - Q_1$. (formula 5.2)

Abaterea cuartilă poate fi absolută – cazul prezentat – sau relativă, raportat la mediană (abaterea intercuartilă relativă).

$$R_{Q\%} = \frac{R_Q}{Q_2} = \frac{Q_3 - Q_1}{Q_2} \quad \text{(formula 5.3)}$$

Dacă împărțim abaterea cuartilă la valoarea 2, adică dacă facem media diferenței dintre Q_3 şi Q_1, obținem **abaterea semi-intercuartilă**.

$$R_{SQ}= R_Q:2 \quad R_{SQ} = \frac{R_Q}{2} = \frac{Q_3 - Q_1}{2} \quad \text{(formula 5.4)}$$

Într-o distribuție de date perfect simetrică, abaterea semi-intercuartilă va fi egală cu mediana. Aceasta reprezintă un indicator al distanței unui scor considerat reprezentativ în raport cu întreaga distribuție.

Dacă, reluăm următoarele date 10, 22, 31, 9, 24, 27, 29, 9, 23, 12 şi apoi le ordonăm, obținem: *9, 9, 10, 12, 22, 23, 24, 27, 29, 31*. Procedura de calcul a cuartilelor este similară medianei. Primul cuartil va fi situat, evident, în poziția 2,75 $Q_1 = \frac{n+1}{4} = \frac{11}{4} = 2{,}75$ deoarece am împărțit şirul ordonat de date în patru părți egale. Urmărind şirul ordonat de date, poziția 2,75 se află între 9 şi 10, adică la 9,5. Al doilea cuartil, mediana (Q_2), se va afla la de două ori poziția primului cuartil, adică 2x2,75 = 5,5. Ați observat probabil că ați obținut exact poziția medianei. În cazul nostru, poziția 5,5 este cuprinsă între 22 şi 23, adică 22,5. Similar, al treilea cuartil (Q_3) este de trei ori poziția primului cuartil, adică 3x2,75=8,25. Iată că, valoarea lui Q_3 se află între 27 şi 29, adică 28. Abaterea cuartilă devine aşadar, $R_Q=Q_3–Q_1=28–9{,}5=18{,}5$. Cu alte cuvinte, între valoarea obținută de 75% din populație şi valoarea obținută

de 25% din populație pot exista cel mult 18,5 valori diferite. Iată că doar 18,5 puncte fac diferența dintre persoanele considerate „slabe” sau cu scoruri inferioare şi persoanele considerate „bune”, cu scoruri superioare. Observați că acest indicator este mai precis în comparație cu amplitudinea împrăştierii. Dacă veți include acum acel scor de 120, veți constata că nu influențează deloc valoarea abaterii cuartile. Abaterea semi-intercuartilă va fi, prin urmare, $R_{SQ}=R_Q:2=18,5:2=9,25$. Acest lucru înseamnă că, teoretic, 9,25 valori ar trebui să fie mai mici decât mediana şi tot atâtea valori mai mari decât mediana. Din grupul oamenilor „medii”, 9,25 valori îi diferențiază pe cei „mediu-slabi” şi 9,25 valori pe cei „mediu-buni”.

Similar medianei, cuartilele pot fi calculate folosind şi date grupate. Dacă vom considera distribuția statistică din tabelul alăturat:

Tabelul 5.1 – Calculul cuartilelor

Clasa	**Distribuția în linii**	**f**	**f_c**
<= 4	||	2	2
5 – 9	|||||	6	8
10 – 14	|||||||	8	16
15 – 19	||||||||||	12	28
20 – 29	||||||||	10	38
30 – 39	||||	4	42
>= 40	||	2	44
		n=44	

Cuartilul inferior (Q_1) va fi situat în poziția $\frac{n+1}{4}=\frac{45}{4}=11,25$, iar quartilul superior (Q_3) în poziția $3Q_1=3x11,25=33,75$. Aşadar, primul cuartil se va afla undeva în intervalul în care se găseşte a 11-a măsurătoare iar al treilea cuartil în intervalul în care se găseşte a 33-a măsurătoare.

În cazul nostru, Q_1 se va găsi în intervalul 10-14 iar Q_3 în intervalul 20-29. Sper că vă mai aduceți aminte formula de calcul a medianei pentru date grupate.

$$Me = l + i\left(\frac{Poz_{Me} - fc_{Me-1}}{f}\right)$$

Dacă ne reamintim, poziția medianei a fost reprezentată prin $\frac{n+1}{2}$ în cazul unui şir de date par sau prin $\frac{n}{2}$ pentru şirul impar de date. Dar cum mediana nu era altceva decât Q_2, valoarea preciza, de fapt, poziția cuartilului al doilea. Deci, singura modificare în formulă ar fi înlocuirea poziției medianei cu poziția cuartilului care ne interesează. Formula generalizată devine cea alăturată, unde P_{Qn} înseamnă poziția cuartilului *n* (1, 2 sau 3) iar fc_{Qn-1} frecvența cumulată a intervalului pre-cuartil de ordin *n*.

$$Qn = l + i\left(\frac{Poz_{Qn} - fc_{Qn-1}}{f}\right) \quad \text{(formula 5.5)}$$

Am specificat mai sus că poziția lui Q_1 este $P_{Q1} = \frac{n+1}{4}$, poziția lui Q_2 este $P_{Q2} = 2 \times \frac{n+1}{4}$ iar poziția lui Q_3 este $P_{Q3} = 3 \times \frac{n+1}{4}$. Aplicând în formule, obținem valorile corespunzătoare cuartilului 1 şi cuartilului 3, astfel:

$$Q1 = l + i\left(\frac{Poz_{Q_1} - fc_{Q1-1}}{f}\right) = 10 + 5\left(\frac{11{,}25-8}{8}\right) = 10 + 5\,\frac{3{,}25}{8} = 12{,}03$$

$$Q3 = l + i\left(\frac{Poz_{Q_3} - fc_{Q3-1}}{f}\right) = 20 + 10\left(\frac{33{,}75-28}{10}\right) = 20 + 10\,\frac{5{,}75}{10} = 25{,}75$$

Abaterea cuartilă va deveni $R_Q = Q_3 - Q_1 = 25{,}75 - 12{,}03 = 13{,}72$ iar abaterea semi-interquartilă $R_{SE} = R_Q : 2 = 13{,}72 : 2 = 6{,}86$.

Abaterea cuartilă este un indicator mai precis decât amplitudinea de variație. Totuşi, dacă distribuția este puternic asimetrică (dacă, de exemplu, foarte multe valori se concentrează în zona cuartilului superior), atunci acest indicator nu furnizează o imagine reală a dispersiei. Se preferă, în acest caz, utilizarea abaterii semi-intercuartile în locul abaterii cuartile, indicator care ne dă numărul de valori diferite ce separă 50% din mijlocul eşantionului însă nu explică modul în care se grupează aceste valori în jurul medianei.

Discutând despre mediană şi cuartile, am văzut că acestea nu reprezintă decât valori la care se situează un anumit procent din eşantion, în condi-

țiile în care setul de date este unul ordonat. Mediana este valoarea la care se situează 50% din eşantion, cuartilele sunt valorile la care se situează 25%, 50%, 75% din eşantion şi aşa mai departe. În termeni generali, putem afla orice valoare la care se situează un anumit procent din eşantion. Astfel, dacă vom dori să aflăm valorile la care se află procente de 20% din eşantion, vom avea cinci categorii – cvintile. Procentele de 33,33% reprezintă valorile trecile şi aşa mai departe. Generic, valoarea la care se află un anumit procent din populație poartă numele de **percentil**. Înțelegerea conceptului de percentile are o importanță deosebită în problematica realizării etaloanelor, normelor, prin metoda cuantilării, metodă folosită de obicei într-o distribuție care nu este normală. Despre aceste lucruri vom discuta însă atunci când vom aborda sistemele de etalonare.

V.2 Indicatori sintetici ai împrăştierii

Indicatorii sintetici ai împrăştierii nu țin cont doar de câteva valori, ca în cazul celor elementari, ci iau în considerare toate măsurătorile care apar în colecția de date şi de aceea au o valoare informativă mult mai mare în comparație cu indicatorii elementari.

V.2.1 Diferența medie Gini

Este un indicator independent de ordinea de efectuare a măsurătorilor iar valoarea acestuia nu depinde de medie. De aceea, diferența medie Gini se foloseşte în distribuții în care media nu este un bun indicator al tendinței centrale (vă mai amintiți probabil de distribuțiile cu scoruri extreme). Indicatorul nu exprimă altceva decât modul în care valorile se împrăştie, în interiorul seriei, unele față de altele. Practic nu este altceva decât o medie aritmetică a diferențelor obținute prin asocierea în perechi, două câte două, a tuturor valorilor din şirul de date. Aceste diferențe se vor lua în valori absolute.

Dacă avem, spre exemplu, următoarele 4 măsurători: 25, 23, 21, 27, atunci, cu aceste valori putem forma un număr de 4^2 adică 16 perechi de diferenţe după cum urmează:

25–25; 25–23; 25–21; 25–27; 23–25; 23–23; 23–21; 23–27; 21–25; 21–23; 21–21; 21–27; 27–25; 27–23; 27–21; 27–27.

Făcând apoi diferenţele, obţinem:

0; 2; 4; -2; -2; 0; 2; -4; -4; -2; 0; -6; 2; 4; 6; 0

Analizând aceste diferenţe, observăm că nu reprezintă altceva decât modul în care se abat valorile unele faţă de altele. Făcând apoi suma acestor valori, luate fără semn, obţinem:

0+2+4+2+2+0+2+4+4+2+0+6+2+4+6+0 = 40

Indicatorul diferenţelor medii a lui Gini nu este altceva decât media diferenţelor luate în valori absolute. Această sumă va fi, aşadar împărţită la numărul de perechi rezultate (16). Obţinem, aşadar, 40:16 = 2,5 care reprezintă valoarea acestui indicator. Mai condensat, am putea scrie acest lucru astfel:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{n} \quad \text{(formula 5.6)}$$

Deşi procedura de calcul a acestui indicator este extrem de simplă, la un număr mare de valori calcularea manuală a diferenţei medii Gini devine imposibilă iar pentru obţinerea indicatorului este absolut necesară utilizarea unui program computer. Diferenţa medie Gini este un indicator „slab" al gradului de împrăştiere şi arată cum se abat, în medie, scorurile unul faţă de celălalt.

V.2.2 Oscilaţia

Calculul oscilaţiei reprezintă tot o medie a diferenţelor dintre perechi, însă, de această dată, ordinea de efectuare a măsurătorilor contează. În mod

normal, sunt rare situațiile în care ne interesează ordinea de efectuare a măsurătorilor. Vizăm acest fapt îndeosebi în experimente, atunci când între măsurători intervenim prin manipularea unor variabile. Dacă avem, de exemplu, următoarele 4 măsurători ale ritmului cardiac al unui subiect la prezentarea unor imagini 80, 85, 91, 96, constatăm că de la o măsurătoare la alta ritmul cardiac s-a accelerat şi putem concluziona că imaginile au avut ca efect creşterea constantă a tensiunii psihice manifestată prin astfel de reacții. Dacă însă ordinea măsurătorilor ar fi fost 80, 96, 91, 85, atunci am observa că avem de a face cu o creştere rapidă a tensiunii psihice urmată de o diminuare progresivă a acesteia. Este cu totul altceva. Iată că, uneori, ordinea măsurătorilor poate avea o semnificație importantă. În asemenea situații se calculează oscilația ca indicator sensibil la aceste diferențe.

Oscilația nu este altceva decât media diferențelor dintre o valoare şi ce anterioară şi se poate calcula în valori absolute sau relative iar numărul acestor diferențe este *n-1*.

Oscilația absolută reprezintă un indicator care furnizează volumul oscilațiilor apărute de la o măsurătoare la alta şi se calculează utilizându-se valorile absolute ale diferențelor după formula următoare:

$$O_{abs} = \frac{\sum_{i=2}^{n} |x_i - x_{i-1}|}{n-1} \quad \text{(formula 5.7)}$$

Reluând cele două exemple anterioare, obținem pentru oscilația absolută a primului set de date, valoarea:

$$O_{abs} = \frac{|85 - 80| + |91 - 85| + |96 - 91|}{3} = 5{,}33$$

Pentru al doilea set de date, obținem, după efectuarea calculelor, următoarea valoare:

$$O_{abs} = \frac{|96 - 80| + |91 - 96| + |85 - 91|}{3} = 9$$

În mod evident cele două valori diferă deoarece ordinea de efectuare a măsurătorilor diferă şi ea. Prima valoare, mai mică, indică un volum mai mic al oscilațiilor, o creştere constantă, deşi rapidă a ritmului cardiac. Valoarea a doua, mare, arată un volum mare al oscilațiilor, o fluctuație a măsurătorilor efectuate.

Oscilația relativă ține cont de semnul diferențelor, spre deosebire de oscilația absolută care lua diferențele în valori fără semn. De aceea, oscilația relativă, deşi nu oferă o imagine a volumului oscilațiilor, ne informează însă asupra direcției lor, în ce sens oscilează măsurătorile, aspect care nu poate fi sesizat de oscilația absolută. De aceea, aceşti doi indicatori se calculează în general împreună. Oscilația relativă are o formulă de calcul asemănătoare celei absolute, ignorându-se modulul:

$$O_{abs} = \frac{\sum_{i=2}^{n}(x_i - x_{i-1})}{n-1} \quad \text{(formula 5.8)}$$

Pentru primul set de date, oscilația relativă rămâne 5,33, în timp ce oscilația relativă pentru al doilea set de date devine 1,66. Ştim că valoarea nu este reprezentativă pentru volumul oscilațiilor, însă comparând cele două valori putem trage concluzii importante. Iată că, la primul subiect, sub influența imaginilor, ritmul său cardiac creşte în mod constant pe măsură ce-i sunt prezentate noi informații, în timp ce la al doilea subiect, după o creştere inițială foarte mare (ştim acest lucru din oscilația absolută) urmează o scădere a ritmului cardiac, ca urmare, probabil, a instalării obişnuinței.

Indicatorii oscilației ne furnizează informații valoroase asupra tendinței tipice sau atipice de evoluție a scorurilor, reprezentând un rezultat important atunci când urmărim să vedem evoluția unui fenomen în timp sau sub influența unor modificări experimentale.

V.2.3 Abaterea medie

Se mai numeşte variație medie, deviație medie, abatere absolută medie sau abatere liniară medie şi se poate calcula atunci când media este un bun indicator al tendinței centrale. Acest indicator informează asupra modului în care se abat, în valori absolute, rezultatele de la medie, acordând aceeaşi pondere tuturor variabilelor. Abaterea medie se poate calcula pe date grupate sau pe date discrete.

Să ne amintim de exemplul cu prietenii. Am văzut că simpla adunare a abaterilor fiecărui scor de la medie nu rezolvă problema verificării gradului de potrivire a modelului statistic cu modelul real, suma lor fiind întotdeauna nulă. Ce-ar fi însă dacă am lua aceste valori în modul, adică nu am ține cont de semn. Am obține, atunci, următoarea valoare:

1,6+0,6+0,4+0,4+1,4=4,4

Rezultă un indicator mai bun al erorii totale, care indică faptul că toate abaterile de la medie sunt în valoare de 4,4, adică numărul de frați cu care se abat toți cei 5 prieteni ai mei de la media de 2,6 este de 4,4 frați. Ideea nu este rea, însă această sumă depinde de numărul de măsurători şi creşte pe măsură ce creşte numărul de cazuri. Dacă aş avea 1000 de prieteni, suma ar avea valori foarte mari. De aceea, se împarte această sumă la numărul de cazuri şi se obține media cu care modelul statistic se abate de la modelul real. Aceasta este şi ideea abaterii medii. În cazul nostru, împărțind 4,4 la 5 obținem 0,88, adică numărul de frați ai prietenilor mei se abate față de medie cu 0,88 frați.

Calculul abaterii medii pe date discrete se realizează după formula următoare, în care x_i reprezintă valoarea măsurătorii, *x barat* reprezintă media iar *n* reprezintă numărul de măsurători.

$$d = \frac{\sum_{i=1}^{n} |x_i - \overline{x}|}{n} \quad \text{(formula 5.9)}$$

Dacă avem următoarele date 10, 22, 31, 9, 24, 27, 29, 9, 23, 12 pentru care media este 19,6 iar numărul de măsurători este 10, avem:

Tabel 5.2 – Calculul abaterii medii

x_i	x_i-m	$\lvert x_i$-m$\rvert$
10	-9,6	9,6
22	2,4	2,4
31	11,4	11,4
9	-10,6	10,6
24	4,4	4,4
27	7,4	7,4
29	9,4	9,4
9	-10,6	10,6
23	3,4	3,4
12	-7,6	7,6
Σ		**76,8**

Practic, formula ne spune să scădem din fiecare valoare valoarea mediei, adică să realizăm diferența dintre scor şi medie. Această diferență o vom trata în termeni de valori absolute (fără semn) după care facem suma acestor valori absolute iar suma o împărțim la numărul de măsurători. Rezultatul obținut reprezintă tocmai valoarea abaterii medii.

În cazul nostru, efectuând calculele conform tabelului alăturat, am obținut suma diferențelor absolute de 76,8. Dacă împărțim această valoare la numărul de măsurători, 10 în cazul nostru, obținem 7,68, valoarea abaterii medii. Deci, pentru date discrete, *d=7,68.*

Pentru date grupate în intervale, vom utiliza în vederea calculării abaterii medii următoarea formulă, unde x_i reprezintă mijlocul intervalului de distribuție, f_i frecvența absolută a acestui interval, *x barat* media şi *n* numărul de măsurători.

$$d = \frac{\sum_{i=1}^{n} |x_i - \overline{x}| \times f_i}{n} \qquad \text{(formula 5.10)}$$

Tabel 5.3 – Calculul abaterii medii (clase)

Clasa	x_i	**f**	$\lvert x_i$-m$\rvert$	$\lvert x_i$-m$\rvert$*f
171 – 174	172,5	9	7,4	66,6
175 – 178	176,5	5	3,4	17
179 – 182	180,5	6	0,6	3,6
183 – 186	184,5	4	4,6	18,4
187 – 190	188,5	3	8,6	25,8
191 -193	192	3	12,1	36,3
		n=30		**Σ =167,7**

Să luăm următorul exemplu de distribuție în clase. Ştim dintr-un capitol anterior că media calculată a acestei distribuții este 179,9. Similar exemplului anterior va trebui să calculăm diferența dintre mijlocul fiecărui interval şi medie, diferență pe care o vom considera în valori absolute. Cu alte cuvinte, dorim să aflăm în ce măsură se aba-

te fiecare interval de la media generală a eşantionului. Urmează apoi să înmulţim aceste abateri cu frecvenţa absolută a fiecărui interval şi să realizăm suma acestor rezultate. În cazul nostru, suma abaterilor intervalelor de la media generală este 167,7. Împărţind această valoare la numărul de măsurători, obţinem 5,59 care nu reprezintă altceva decât abaterea medie *(d)* calculată pe intervale de clase.

V.2.4 Abaterea mediană

Se mai numeşte şi variaţie mediană sau deviaţie mediană şi se utilizează în situaţia în care nu putem folosi media, fie datorită faptului că nivelul de măsură al variabilei este unul ordinal, fie că avem scoruri extreme într-o distribuţie, fapt ce face media un indicator inadecvat.

Modalitatea de calcul a abaterii mediane este similară celei de mai sus (abaterea medie), sigura diferenţă fiind aceea că, în loc de valoarea mediei, vom folosi valoarea medianei. Calculul acesteia vi-l lăsăm dumneavoastră ca exerciţiu.

***Exerciţiu**:*

Calculaţi abaterea mediană pentru datele discrete şi datele grupate în intervale de clasă utilizând exemplele de mai sus.

V.2.5 Dispersia

Atunci când putem utiliza media, un indicator frecvent este indicatorul dispersiei sau indicatorul varianţei (atenţie, nu variaţie ci varianţă) notat cu sigma pătrat pentru populaţie sau s^2 în cazul unui eşantion. Varianţa reprezintă media pătratică a diferenţelor valorilor individuale faţă de media aritmetică şi are o relaţie directă cu cel mai important indicator al gradului de împrăştiere a rezultatelor, abaterea standard.

În loc să folosim modul pentru a realiza suma diferenţelor, preferăm să ridicăm la pătrat aceste diferenţe. Ştim foarte bine că dacă ridicăm un nu-

măr cu acelaşi semn la pătrat, rezultatul este întotdeauna pozitiv. Iată o altă modalitate prin care evităm anulările reciproce determinate de semn.

Pentru fraţii prietenilor mei, vom avea:

$$(-1,6)^2+(-0,6)^2+(0,4)^2+(0,4)^2+(1,4)^2$$

$$2,56+0,36+0,16+0,16+1.96 = 5,20$$

Această valoare, 5,20 poartă numele de **suma abaterilor pătratice** şi reprezintă un indicator mai bun decât suma abaterilor simple luate în modul. La fel, această sumă depinde de numărul de cazuri şi de aceea, pentru a compensa acest lucru, se împarte suma abaterilor pătratice la numărul de cazuri. În situaţia noastră, 5,20:5=1,04, exact valoarea varianţei.

Ajunşi aici, se impune totuşi o precizare. Împărţirea la numărul cazurilor (*n*) se realizează atunci când vorbim de o populaţie. Dacă ne referim la eşantion, atunci nu împărţim la numărul de cazuri ci la numărul de cazuri minus unu (*n-1*). Aceasta este o ajustare statistică, iar motivele care o susţin vor fi explicate mai târziu. Iată că, 5,20:4 = 1,30, adevărata măsură a varianţei.

Formula de calcul a dispersiei este următoarea, în care x_i reprezintă valoarea măsurată, *x barat* media iar *n* numărul de măsurători.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n} \qquad \text{(formula 5.11)}$$

În activitatea curentă, deoarece lucrăm cu eşantioane şi nu cu populaţii, se înlocuieşte *n* cu *n-1*, astfel încât formula dispersiei pentru eşantioane devine:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} \qquad \text{(formula 5.12)}$$

Şi dispersia, se poate calcula pe date discrete sau pe date grupate în intervale.

Reluând exemplul anterior, ne amintim că avem următoarele date 10, 22, 31, 9, 24, 27, 29, 9, 23, 12 pentru care media este 19,6 iar numărul de măsurători este 10. Vom utiliza un tabel similar cu cel prezentat în exemplul de mai sus şi vom analiza formula.

Tabel 5.4 – Calculul dispersiei

x_i	x_i-m	$(x_i$-m$)^2$
10	-9,6	92,16
22	2,4	5,76
31	11,4	129,96
9	-10,6	112,36
24	4,4	19,36
27	7,4	54,76
29	9,4	88,36
9	-10,6	112,36
23	3,4	11,56
12	-7,6	57,76
Σ		**684,4**

Privind cu atenție, observăm că trebuie să facem, prima dată, diferența dintre scorul brut şi medie (x_i-m), la fel ca în procedura de calcul a abaterii medii. Deoarece obținem valori negative şi valori pozitive, dacă le-am însuma în această etapă, valorile contrare se vor anula reciproc şi nu am obține nici o informație sintetică utilă. Pentru a preveni acest lucru, vom ridica la pătrat fiecare diferență. În acest fel compensăm anularea reciprocă ce poate să apară la însumarea valorilor, toate valorile rezultate fiind pozitive. După ridicarea la pătrat, putem face fără să ne preocupe anularea valorilor, suma rezultatelor (suma abaterilor pătratice). În cazul nostru, suma abaterilor pătratice este 684,4. Împărțind acum suma pătratelor la numărul de măsurători (10), obținem valoarea 68,44 ca valoare a indicatorului varianței sau dispersiei. Deoarece lucram cu un eşantion şi nu la nivelul populației, împărțirea o vom face la *n-1* (10-1=9) rezultând valoarea corectată a dispersiei. Deci, pentru setul nostru de date, varianța s^2=70,04

Pentru date grupate în intervale de clasă, formula de calcul a varianței se modifică puțin, unde f_i reprezintă frecvența absolută a intervalului de clasă, x_i reprezintă mijlocul acestui interval, restul elementelor fiind cunoscute.

$$\sigma^2 = \frac{\sum_{i=1}^{n} f_i \times (x_i - \overline{x})^2}{n} \quad \text{(figura 5.13)}$$

$$s^2 = \frac{\sum_{i=1}^{n} f_i \times (x_i - \overline{x})^2}{n-1} \quad \text{(figura 5.14)}$$

Reluând distribuţia prezentată în secţiunea anterioară, ne reamintim că media acestor rezultate a fost de 179,91 iar numărul de cazuri 30.

Primul lucru pe care îl vom face va fi să calculăm diferenţa dintre mijlocul fiecărui interval de clasă şi medie. Din motivele expuse anterior, nu putem face în acest moment suma diferenţelor obţinute. Va trebui să ridicăm la pătrat aceste diferenţe după care să înmulţim pătratele obţinute cu frecvenţa absolută a fiecărui interval. După această operaţiune, putem calcula suma abaterilor pătratice, care, în cazul nostru, va deveni 1298,55. Împărţind această sumă la numărul de măsurători (n=30) obţinem valoarea varianţei 42,28. Totuşi, numărul de măsurători fiind mic, acest rezultat în corectăm împărţind nu la 30 ci la 29. Va rezulta astfel 44,77, valoarea corectată a varianţei.

Tabel 5.5 – Calculul dispersiei (clase)

Clasa	**x_i**	**f**	**x_i-m**	**$(x_i-m)^2$**	**$f*(x_i-m)^2$**
171 – 174	172,5	9	7,4	54,76	492,84
175 – 178	176,5	5	3,4	11,56	57,8
179 – 182	180,5	6	0,6	0,36	2,16
183 – 186	184,5	4	4,6	21,16	84,64
187 – 190	188,5	3	8,6	73,96	221,88
191 -193	192	3	12,1	146,41	439,23
		n=30			**Σ =1298,55**

V.2.6 Abaterea standard

Se notează cu *s* în cazul unui eşantion sau cu *sigma* în cazul unei populaţii şi o putem întâlni sub numele de abatere etalon, abatere tip, abatere pătratică medie. Acest indicator este cel mai precis şi inteligibil pentru măsura gradului de împrăştiere al rezultatelor în jurul tendinţei centrale şi contribuie la definirea distribuţiei normale. Poate fi calculat şi dobândeşte semnificaţie numai atunci când poate fi calculată media.

Varianţa ne dă expresia medie a abaterilor dintre medie şi observaţii, însă această valoare este exprimată sub formă ridicată la pătrat. Nu ştiu dacă vă este foarte utilă o expresie de genul „în medie, numărul de fraţi ridicat la pătrat cu care se abat fraţii prietenilor mei de la media de 2,6 fraţi este de 1,3

fraţi". Cred că mai degrabă vă poate crea confuzie. Acest lucru se întâmplă deoarece, pentru a evita problemele legate de semn, am ridicat la pătrat diferenţele. Cum scăpăm acum de această ridicare la pătrat? Vă mai amintiţi probabil că inversul ridicării la pătrat este extragerea rădăcinii pătrate (a radicalului). Exact acest lucru îl vom face şi noi. Extrăgând radical din 1,3 vom obţine valoarea 1,14, expresia abaterii standard. Putem acum afirma că, în medie, numărul de fraţi ai prietenilor mei se abate de la valoarea de 2,6 cu 1,14.

Abaterea standard se calculează după formula următoare, unde toate datele ne sunt deja cunoscute. Dacă privim cu atenţie această formulă, constatăm că expresia de sub radical nu este altceva decât varianţa. Prin urmare, după ce calculăm varianţa, putem afla uşor abaterea standard extrăgând radicalul de ordin doi din varianţă.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i \times (x_i - \overline{x})^2}{n}} = \sqrt{\sigma^2} \quad \text{(formula 5.15)}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i \times (x_i - \overline{x})^2}{n-1}} = \sqrt{s^2} \quad \text{(formula 5.16)}$$

În primul exemplu, cel în care am folosit date discrete, varianţa obţinută a fost s^2=68,44. Deci, abaterea standard va fi egală cu rădăcina pătrată din 68,44, adică 8,27. Similar, pentru datele grupate în intervale, varianţa a fost 44,77 iar abaterea standard devine 6,69.

Abaterea standard măsoară modul în care rezultatele se împrăştie în jurul mediei, valorile mari indicând un grad mare de împrăştiere. Este, dacă doriţi, un indicator ce arată în ce măsură media este reprezentativă pentru observaţii, în ce măsură modelul statistic se potriveşte cu modelul real. Abaterile standard mici arată că valorile sunt puţin distanţate faţă de medie, acest indicator al tendinţei centrale reprezentând foarte bine datele. Valorile mari ale abaterii standard arată că datele sunt distanţate, împrăştiate puternic în

jurul mediei, aceasta nefiind un bun model statistic. Abaterea standard zero indică faptul că avem aceleaşi valori (de exemplu o distribuție 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5 are media 5 şi abaterea standard 0).

Comparând cele două rezultate de mai sus, constatăm că în primul exemplu rezultatele sunt mai împrăştiate în jurul mediei decât în al doilea exemplu. Cu cât abaterea standard este mai mică, cu atât datele sunt mai grupate în jurul tendinței centrale.

Abaterea standard se exprimă în aceleaşi unități de măsură ca şi media. Dacă media înălțimii unor subiecți este de 179,91 centimetri, atunci abaterea standard de 6,69 va fi exprimată tot în centimetri. Convenim aşadar să descriem o distribuție de acest gen prin media înălțimii subiecților din eşantion egală cu 179,91 centimetri cu o abatere standard de 6,69 centimetri.

Ridicarea la pătrat din formula dispersiei, introduce o oarecare distorsiune în analiza împrăştierii datelor, deoarece prin această operație se atribuie o pondere mai mare valorilor aflate la o distanță mare față de medie.

Rețineți că împărțirea la numărul de cazuri *n* este valabilă pentru o variabilă teoretică sau în cazul populației. Deoarece, în general, lucrăm cu eşantioane şi nu cu populații, vom folosi în formulă împărțirea la *n-1*, valabilă pentru variabila empirică. Această corecție se foloseşte în vederea evitării subestimării gradului de împrăştiere a rezultatelor în jurul mediei şi poartă numele de **grade de libertate.**

Media şi abaterea standard sunt cei doi indicatori importanți în analiza distribuției normale. Utilizarea lor este condiționată de absența unor scoruri extreme şi de variabile situate cel puțin la un nivel de măsură de interval.

La fel ca şi media, indicatorul principal al tendinței centrale, abaterea standard are şi ea o serie de proprietăți utile în practică:

- Dacă adăugăm sau scădem o valoare constantă la fiecare scor din distribuție, abaterea standard nu suferă modificări;

- Dacă înmulțim sau împărțim fiecare scor din distribuție cu o constantă, abaterea standard se multiplică sau se divide cu valoarea acelei constante;
- Distanța între valoarea abaterii standard şi medie este mai mică în comparație cu distanța dintre abaterea standard şi orice alt scor. Cu alte cuvinte, media se distanțează cel mai puțin de abaterea standard în comparație cu orice alt scor.

V.2.7 Coeficientul de variație

Prezintă, mai intuitiv decât abaterea standard, gradul de împrăştiere al rezultatelor în jurul mediei, deoarece este o expresie procentuală a împărțirii abaterii standard la medie. De asemenea, pentru datele ordinale, se poate calcula coeficientul de variabilitate ca expresie procentuală a împărțirii abaterii mediane la mediană. Formulele de calcul în acest caz sunt:

$$v_x = \frac{s}{\bar{x}} \times 100, \text{în } cazul\ mediei \qquad \text{(formula 5.17)}$$

$$v_{Md} = \frac{d_{Md}}{Md} \times 100, \text{în } cazul\ medianei \qquad \text{(formula 5.18)}$$

Prin înmulțirea coeficientului de variație cu 100 obținem exprimarea procentuală a acestui indicator, mai intuitivă în comparație cu valoarea brută.

În exemplul nostru, *v = 7,04:19,6=0,35x100=35*, adică o variație de 35% în cazul datelor discrete şi *v = 6,69:179,91=0,03x100=3*, o variație de doar 3,00% pentru datele grupate în interval.

Cu cât valoarea acestui coeficient este mai mare, cu atât media este un indicator mai imprecis pentru tendința centrală. Valoarea acestui coeficient nu depinde de unitatea de măsură, ca în cazul abaterii standard şi de aceea poate fi folosit în situații în care comparăm, sub aspectul variabilității, două distribuții diferite.

Există anumite reguli ce permit aprecierea gradului de împrăştiere a scorurilor în jurul mediei utilizându-se acest coeficient (Popa, 2006):

- Un coeficient de variaţie mai mic de 15% indică un grad de împrăştiere redus, media fiind un bun indicator al tendinţei centrale;
- Un coeficient de variaţie cuprins între 15% şi 30% arată o împrăştiere moderată, caz în care reprezentativitatea mediei este satisfăcătoare;
- Un coeficient de variaţie de peste 30% arată o împrăştiere mare, media nemaifiind un bun indicator pentru tendinţa centrală.

V.3 Obţinerea indicatorilor împrăştierii în SPSS

Pentru a studia modul în care aflăm indicatorii dispersiei în SPSS, vom folosi baza de date *„IQ.sav”*, în care avem deja datele introduse în capitolul anterior. Folosim, pentru aceasta, tot comanda **„Analyze”**, apoi submeniul **„Descriptive Statistics”** şi opţiunea **„Frequencies…”**. Se va deschide binecunoscuta de acum fereastră, pe care am studiat-o în momentul în care am aflat cum putem calcula indicatorii tendinţei centrale. Şi de data aceasta vom folosi, ca exemplu, variabila „Varsta subiectilor”, variabilă pe care o vom transfera din lista variabilelor din baza de date în lista variabilelor supuse analizei (**„Variable(s)”**) din partea dreaptă a formularului, aşa cum se poate observa în figura 5.4. Asupra modu-

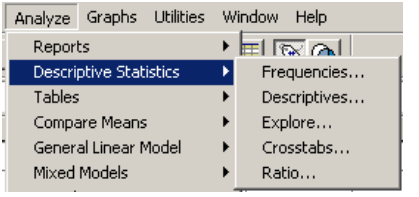


Figura 5.3 – Meniul de analiză a frecvenţelor

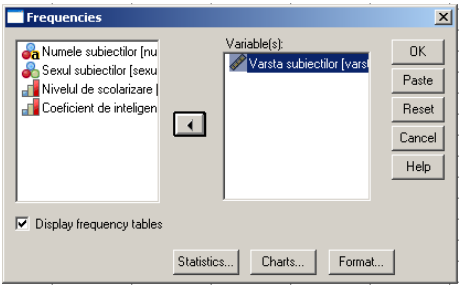


Figura 5.4 – Includerea variabilei în analiză

lui în care se face acest transfer nu mai insistăm, deoarece procedeul a fost descris pe larg în capitolul destinat indicatorilor tendinței centrale. Dacă nu vă puteți aminti, revedeți capitolul 3.

După transfer, va trebui să comunicăm aplicației SPSS să ne calculeze indicatorii împrăştierii. Vă amintiți probabil că acest lucru îl putem face apăsând butonul **„Statistics…"** situat în partea de jos a acestui formular. Imediat, se va deschide o nouă fereastră, cunoscută de la calculul indicatorilor tendinței centrale. Observăm, în partea din stânga-jos a acestei ferestre, secțiunea destinată calculului indicatorilor dispersiei (secțiunea **„Dispersion"**). Vom putea calcula abaterea standard, prin bifarea casetei **Std. deviation"**, varianța, bifând caseta „**Variance"** sau amplitudinea de variație cu ajutorul casetei „**Range"**. De asemenea este posibilă afişarea valorii minime şi maxime din distribuția noastră prin bifarea casetelor **„Minimum"** şi **„Maximum"**. Secțiunea permite şi calcularea erorii standard a mediei (caseta „**S.E mean"**) însă asupra acestui element vom reveni ceva mai târziu.

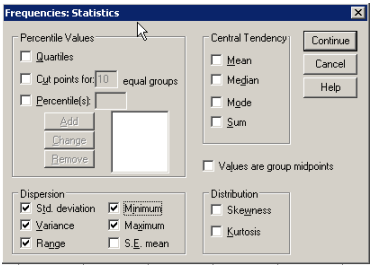


Figura 5.5 – Selectarea indicatorilor împrăştierii

Vom alege, după cum puteți observa în figura de mai sus, afişarea abaterii standard, a varianței, amplitudinii distribuției, minimului şi maximului. Apăsăm apoi butonul **„Continue"** pentru ca SPSS să memoreze comenzile noastre şi ne întoarcem în formularul anterior. Nu ne mai rămâne decât să apăsăm butonul „**OK"** pentru a lansa procedeul de analiză.

Tabel 5.6 – Indicatorii împrăştierii

Statistics

Varsta subiectilor

N	Valid	30
	Missing	0
Std. Deviation		5,523
Variance		30,506
Range		19
Minimum		20
Maximum		39

După câteva fracțiuni de secundă SPSS va afişa în fereastra de rezultate tabelul cu rezultatele comandate de noi. Automat, SPSS include şi numărul de cazuri

(*N*). Aflăm, din analiza acestui tabel, că avem un număr de 30 de subiecţi (*N=30*) şi toţi subiecţii au valori pentru variabila „Varsta subiectilor" (**Valid=30; Missing=0**). Cu alte cuvinte, nici unul dintre aceşti subiecţi nu are necompletată valoarea pentru variabila vârstă. Aflăm, în continuare, că abaterea standard (**Std. Deviation**) este 5,523 iar varianţa (**Variance**) este 30,506. Dacă veţi extrage rădăcina pătrată din varianţă (30,506) veţi obţine exact valoarea abaterii standard (5,523), ceea ce confirmă teoria expusă mai sus. De asemenea, amplitudinea de variaţie (**Range**) este 19, cu alte cuvinte între **Minimum** 20 şi **Maximum** 39 pot exista cel mult 19 valori posibile. Pare mult mai simplu decât să efectuăm manual calculele prezentate la începutul acestui capitol. Iată de ce SPSS îşi merită cu prisosinţă banii pe care îi solicită producătorul…

Concluzii:

- *Indicatorii tendinţei centrale nu pot caracteriza singuri o distribuţie. Este necesar să studiem şi modul în care rezultatele se „împrăştie" în jurul tendinţei centrale prin calculul indicatorilor dispersiei;*
- *Indicatorii dispersiei pot ţine seama doar de anumite date şi se numesc indicatori elementari ai împrăştierii sau de toate datele, cazul indicatorilor sintetici;*
- *Indicatorii elementari ai împrăştierii sunt:*
    - *Amplitudinea de variaţie absolută şi relativă;*
    - *Intervalul cuartil şi abaterea cuartilă;*
- *Indicatorii sintetici ai împrăştierii sunt:*
    - *Diferenţa medie Gini;*
    - *Oscilaţia absolută şi relativă;*
    - *Abaterea medie;*
    - *Abaterea mediană;*
    - *Dispersia sau varianţa;*
    - *Abaterea standard;*
    - *Coeficientul de variaţie.*
- *Media şi abaterea standard sunt indicatorii pe baza cărora se caracterizează o distribuţie normală;*
- *Indicatorii dispersiei bazaţi pe medie pot fi calculaţi numai dacă nivelul de măsurare al variabilei este cel puţin unul de interval şi nu există scoruri extreme în distribuţie;*

Întrebări şi exerciţii:

- *Calculaţi abaterea standard şi coeficientul de variaţie al următoarei distribuţii statistice precizând etapele de rezolvare:.*

Clasa	**f**
3 – 5	3
6 – 8	5
9 – 11	9
12 – 14	16
15 – 17	10
18 – 20	4
21 – 23	3
24 – 26	1

- *Calculaţi indicatorii dispersiei în cazul următoarelor seturi de date:*
  - *12, 23, 9, 6, 14, 14, 12, 25, 9, 12*
  - *1, 4, 5, 6, 19, 1, 5, 3, 16, 12, 5, 4*
  - *32, 56, 91, 16,32, 5, 14,62,19, 12*
- *Extrageţi indicatorii tendinţei centrale pentru variabila „iq” folosind SPSS*

VI. DISTRIBUŢIA NORMALĂ

În acest capitol se va discuta despre:

- Distribuţia normală;
- Analiza distribuţiei normale;
- Note standardizate;
- Inventarul statistic de bază
- Etaloane şi sisteme de etalonare.

După parcurgerea capitolului, cursanţii vor fi capabili să:

- Înţeleagă caracteristicile unei distribuţii normale;
- Analizeze o distribuţie sub aspectul caracteristicilor acesteia şi să precizeze dacă este sau nu normală;
- Realizeze şi să interpreteze inventarul statistic de bază utilizând SPSS;
- Transforme o notă brută în note standardizate şi să precizeze poziţia pe care o ocupă scorul particular;
- Decidă ce sistem de etalonare poate fi utilizat în funcţie de distribuţie şi să realizeze un etalon.

Am menţionat deseori acest concept al distribuţiei normale şi a venit momentul să îl abordăm în detaliu. Vom începe, ca de obicei, cu un exemplu: dacă am avea posibilitatea să măsurăm înălţimea tuturor bărbaţilor din România, am observa că există foarte puţini pitici şi foarte puţini giganţi; ceva mai mulţi oameni mici şi, la fel, ceva mai mulţi oameni foarte înalţi iar cei mai mulţi oameni sunt de statură medie. Care ar fi aşadar probabilitatea ca să întâlnim în populaţia masculină un gigant? Foarte mică.

Figura 6.1 – Distribuţia înălţimii tuturor bărbaţilor din România

Ceva mai mare ar fi probabilitatea de a întâlni în populație un om înalt şi foarte mare să întâlnim, de exemplu, un om cu înălțimea de 178 centimetri, înălțime medie. Dacă am reprezenta grafic, printr-un grafic cu bare, înălțimea tuturor bărbaților din România, am obține o figură asemănătoare cu figura de mai sus.

Privind cu atenție figura, observăm că, la înălțimi ale subiecților foarte mici (piticii), frecvența de apariție este şi ea foarte mică, apropiată de zero. Pe măsură ce înălțimea creşte, creşte şi frecvența de apariție a acesteia, astfel încât va atinge un maximum în zona tendinței centrale, după care începe din nou să scadă spre înălțimi mari. La înălțimi foarte mari (giganții), frecvența de apariție se apropie din nou de zero.

Această distribuție este o distribuție normală şi poate fi evaluată numai în cazul unei variabile continue. De aceea, distribuția normală se mai numeşte şi distribuție continuă. O distribuție normală, este pe deplin caracterizată de medie, ca indicator al tendinței centrale şi de abaterea standard, ca indicator al dispersiei. Aceşti doi indicatori poartă numele de **parametri ai repartiției normale**. Dacă cunoaştem media şi abaterea standard, putem oricând calcula probabilitatea de apariție a unei valori particulare în această distribuție.

Distribuția normală, a fost descrisă prima dată de Ch. Fr. Gauss (1777-1855) şi de aceea distribuția normală se mai numeşte şi **distribuție gaussiană**. Deoarece la demonstrarea acestui concept a participat şi P.S. Laplace (1749-1827), în literatura de specialitate se va întâlni şi termenul de distribuție gauss-laplace. Toți aceşti termeni se referă la acelaşi lucru, distribuția normală.

Am văzut deja că media şi abaterea standard descriu complet o distribuție normală. În literatură, se foloseşte pentru medie simbolul *m* sau *x* barat iar pentru abaterea standard simbolul *s,* atunci când vorbim de un eşantion.

La nivel teoretic sau atunci când vorbim de populații, folosim simbolul μ pentru medie şi simbolul σ pentru abaterea standard.

Pentru ca o distribuție să fie considerată normală, vor trebui îndeplinite simultan următoarele condiții:

- **Să fie unimodală** – adică să existe un singur mod, o singură categorie cu frecvență maximă;
- **Să fie simetrică față de medie** – adică să nu fie deplasată spre stânga sau spre dreapta;
- **Să fie normal boltită** – adică să nu fie nici ascuțită (foarte omogenă) şi nici turtită (foarte eterogenă);

De asemenea, limitele din stânga şi din dreapta ale unei distribuții normale tind spre valoarea zero, pe care însă nu o întâlnesc niciodată. O distribuție perfect normală are aceeaşi valoare pentru toți cei trei indicatori ai tendinței centrale (media, mediana şi modul), adică *media = mediana = mod*. În practică, acest lucru se întâlneşte extrem de rar şi, de aceea, ne punem problema între ce limite putem considera o distribuție ca fiind normală.

VI.1 Modalitatea

O distribuție normală este o distribuție unimodală, adică există doar o singură categorie cu frecvența maximă. Prezența a două sau mai multe valori modale determină distribuții bimodale, trimodale, în general polimodale (multimodale), distribuții ce nu pot fi considerate ca fiind distribuții normale.

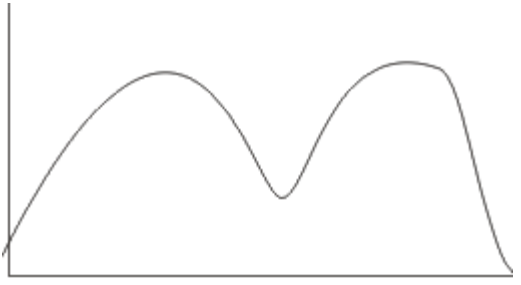


Figura 6.2 – Distribuție bimodală

O distribuție ca cea din figura 6.1, este o distribuție normală. O distribuție similară distribu-

ției din figura 6.2, este o distribuție bimodală, deoarece observăm existența a două categorii cu frecvența egală şi maximă.

Normalitatea distribuției, sub aspectul modalității, se verifică prin calcularea valorii mod. Dacă există o singură valoare cu frecvența maximă, atunci distribuția poate fi considerată normală sub acest aspect. În condițiile în care sunt două sau mai multe valori cu frecvența maximă şi, evident, egală, distribuția este polimodală (multimodală) şi nu poate fi considerată ca fiind normală.

VI.2 Simetria

O distribuție este simetrică, dacă valorile sunt egal (simetric) răspândite în jurul tendinței centrale. Atunci când rezultatele tind către valori mici, sunt aglomerate în partea stângă a distribuției, avem de a face cu o distribuție asimetrică spre dreapta (sau distribuție skewness pozitiv). Când rezultatele tind către valori mari, se aglomerează în partea dreaptă a distribuției, vorbim despre o distribuție asimetrică la stânga (skewness negativ). Iată că, asimetria este dată de panta distribuției şi nu de vârful acesteia, aşa cum tratează şi consideră unii.

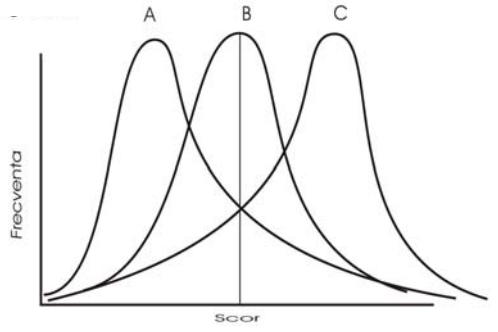


Figura 6.3 – Distribuție simetrică (B), asimetrică la dreapta (A) şi asimetrică la stânga (C)

În figura 6.3, distribuția „B" este o distribuție simetrică. Distribuția „A" este o distribuție asimetrică la dreapta (skewness pozitiv) unde predomină scorurile scăzute, în timp ce distribuția „C" este o distribuție asimetrică la stânga (skewness negativ) în care predomină scorurile mari.

Nu ne putem limita la o apreciere pur vizuală a simetriei; acest lucru necesită demonstrarea statistică a simetriei sau asimetriei distribuției. În acest scop, există mai mulți coeficienți care pot fi calculați.

VI.2.1 Coeficientul Yule

Coeficientul măsoară gradul de asimetrie prin raportul dintre diferența şi suma abaterilor cuartile şi se calculează după formula alăturată. Observăm că, pentru a calcula acest coeficient, este necesar să calculăm pur şi simplu cele trei cuartile, după metoda prezentată într-unul dintre capitolele anterioare.

$$C_Y = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$ (formula 6.1)

Coeficientul poate lua, în general, valori de la -1 la +1 şi indică mărimea asimetriei negative sau pozitive. Cu cât valorile sunt mai apropiate de 0, cu atât distribuția este mai simetrică. Cu cât se apropie de -1 sau de +1, ea devine tot mai asimetrică spre stânga sau spre dreapta.

Acest coeficient este foarte uşor de calculat, însă ține cont doar de câteva valori – valorile cuartile – şi nu de toate valorile din distribuție, fapt care îi limitează precizia şi se consideră că este doar un coeficient elementar al simetriei.

Să ne reamintim exemplul de calcul al cuartilelor, din capitolul anterior. Am stabilit atunci şi nu vom mai intra acum în detalii, că primul cuartil (Q_1) are valoarea 12,03, al doilea cuartil (Q_2) este 16,16 iar al treilea cuartil (Q_3) avea valoarea 25,75. Înlocuind în formula noastră, obținem valoarea negativă -1,35, indicând o asimetrie spre stânga, adică o tendință către scoruri mari, fapt ce se poate observa cu uşurință din distribuția statistică.

VI.2.2 Coeficientul de asimetrie Fisher

Coeficientul de asimetrie al lui Fisher, nu se bazează pe câteva elemente, precum coeficientul Yule, ci pe toate valorile din distribuție şi se cal-

culează folosindu-se momentele centrate. Varianța, de exemplu, este momentul centrat de ordin doi, deoarece am ridicat la pătrat fiecare diferență dintre o valoare particulară şi medie. Dacă ridicăm aceste diferențe la cub, obținem momentul centrat de ordin 3 iar dacă ridicăm la puterea a patra, momentul centrat de ordin patru. Ştim, de la abaterea standard, că ridicarea la pătrat amplifică ponderea pe care o au variabilele distanțate puternic de la medie. Această pondere este amplificată şi mai mult prin ridicare la cub, însă, de această dată, va fi luat în considerare şi semnul cubului diferențelor.

$$\beta_1 = \frac{\mu_3}{\sigma^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n\sigma^3}$$ (formula 6.2)

Dacă raportăm momentul de ordin trei la cubul abaterii standard, obținem exact expresia **coeficientului de asimetrie a lui Fisher**.

Iată o primă formulă care, probabil, vă va determina să aruncați cartea din mână. Dacă ați ajuns la această frază, dacă acum continuați să citiți, înseamnă că dumneavoastră chiar doriți să învățați serios aceste noțiuni… Ideea este că, deşi pare complexă, formula nu face altceva decât să calculeze un fel de … varianță, doar că nu mai ridicăm numai la pătrat fiecare diferență între scorul obținut, ci şi la cub. Procedeul nu este deloc mai complicat decât calculul varianței, după cum vom vedea în următoarele exemple.

Tabel 6.1 – Calculul coeficientului de asimetrie

x_i	x_i-m	$(x_i\text{-m})^2$	$(x_i\text{-m})^3$
10	-9,6	92,16	-884,736
22	2,4	5,76	13,824
31	11,4	129,96	1481,544
9	-10,6	112,36	-1191,02
24	4,4	19,36	85,184
27	7,4	54,76	405,224
29	9,4	88,36	830,584
9	-10,6	112,36	-1191,02
23	3,4	11,56	39,304
12	-7,6	57,76	-438,976
Σ		**684,4**	**-850,08**

Reluând exemplul cu date discrete, ne amintim că avem următoarele date 10, 22, 31, 9, 24, 27, 29, 9, 23, 12 pentru care media este 19,6 iar numărul de măsurători este 10.

De asemenea, am aranjat aceste date într-un tabel pentru a ne uşura procedeul de calcul. Tot la calculul dispersiei, vă amintiți, am făcut pătratul diferențelor $(x_i\text{-}m)^2$, după care suma pătratelor.

În formulă trebuie să mai calculăm cubul acestor diferențe (pentru momentul centrat de ordin trei). Vom mai adăuga o coloană pe care o vom denumi $(x_i\text{-}m)^3$ şi vom ridica la puterea a treia fiecare diferență $x_i\text{-}m$. La final vom face suma pătratelor şi suma cuburilor. Suma cuburilor poate fi un număr pozitiv sau negativ, în funcție de ponderea pe care o au abaterile scorurilor înspre partea dreaptă sau înspre partea stângă față de medie. În acest moment nu ne rămâne decât să înlocuim în formulă. Suma pătratelor este 684,4, suma cuburilor este -850,08 iar numărul de măsurători este 10. Înlocuind, obținem următoarele momente de ordin trei şi doi, prezentate în figura alăturată. După efectuarea calculelor obținem un coeficient de asimetrie de - 0,15, coeficient care indică o distribuție simetrică.

$$\beta_1 = \frac{\mu_3}{\sigma^3} = \frac{-850{,}08}{10 \times \left(\sqrt{\frac{684{,}4}{10}}\right)^3} = \frac{-850{,}08}{5661{,}93} = -0{,}15$$

Similar se realizează calculele şi în cazul datelor grupate în clase, derivând din formula de calcul a dispersiei pentru date grupate. Acest exemplu îl lăsăm ca exercițiu.

Acest coeficient are valoarea zero pentru distribuții perfect simetrice. Valorile negative sunt considerate ca asimetrii la stânga iar valorile pozitive indică asimetrii la dreapta. O valoare zero este însă o valoare pe care o vom întâlni foarte rar în practică. În mod normal, valorile sunt diferite de zero. Nu însă orice valoare diferită de zero poate indica o asimetrie care să devieze semnificativ de la distribuția normală. Vom vedea, puțin mai târziu, care sunt limitele între care putem considera totuşi o distribuție asimetrică fiind normală.

Exercițiu:

Calculați coeficientul de asimetrie folosind distribuția în clase prezentată la capitolul de calcul a dispersiei.

Atunci când discutăm de asimetrie, trebuie să precizăm o serie de relații care apar între indicatorii tendinței centrale, medie, mediană şi mod.

- **Într-o distribuție perfect simetrică**, media mediana şi modul au exact aceleaşi valori. Caracteristica acestei distribuții o reprezintă coincidența absolută a celor trei indicatori ai tendinței centrale. După cum puteți observa în figura alăturată, atât media, cât şi mediana şi modul, se află în acelaşi punct, la mijlocul distribuției. Grupa subiecților cu scoruri mai mici decât media este perfect simetrică cu grupa subiecților cu scoruri mai mari decât media. Acest aspect este menționat din rațiuni teoretice. În practică nu vom întâlni o distribuție perfect normală, ci o distribuție care poate fi acceptată ca fiind o distribuție normală.

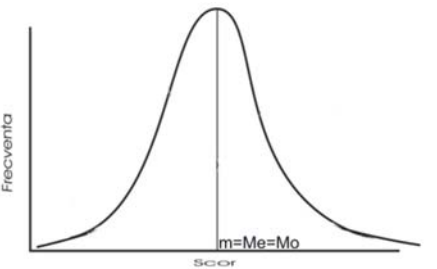


Figura 6.4 – O distribuție perfect simetrică

- **Într-o distribuție asimetrică la dreapta** (distribuție skewness pozitiv, vezi figura alăturată), predomină scorurile mici. În acest caz, modul este valoarea situată cel mai la stânga în şirul de date, iar

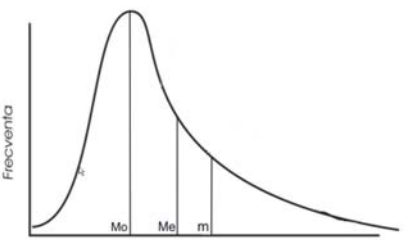


Figura 6.5 – O distribuție asimetrică la dreapta

mediana este mai mică decât media. Evident, mediana fiind valoarea care împarte şirul ordonat de date în două părţi egale, iar dacă în distribuţie predomină scorurile mici, atunci le mari sunt considerate ca scoruri extreme. Ştim, de la analiza preciziei indicatorilor tendinţei centrale, că într-o serie de date în care întâlnim scoruri extreme mari, media tinde să le pună în valoare. Iată că acest fapt este ilustrat grafic în figura de mai sus. Observaţi relaţia existentă într-o asemenea distribuţie: **Mo<Me<m**. Această relaţie este relaţia caracteristică a unei distribuţii asimetrice pozitiv.

- **Într-o distribuţie asimetrică la stânga** (distribuţie skewness negativ, vezi figura alăturată), predomină scorurile mari. În acest caz, modul este valoarea situată cel mai la dreapta în şirul de date, iar mediana este mai mare decât media. Evident, mediana fiind valoarea care împarte şirul ordonat de date în două părţi egale, iar dacă în distribuţie predomină scorurile mari, atunci scorurile mici sunt considerate ca scoruri extreme. Ştim, de la analiza preciziei indicatorilor tendinţei centrale, că într-o serie de date în care întâlnim scoruri extreme mici, media tinde să le pună în valoare. Iată că acest fapt este ilustrat grafic în figura de mai sus. Observaţi relaţia existentă într-o asemenea distribuţie: **Mo>Me>m**. Această relaţie este relaţia caracteristică a unei distribuţii asimetrice negativ.

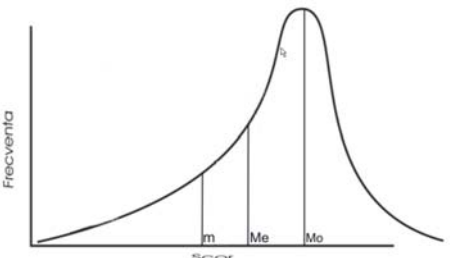


Figura 6.6 – O distribuţie asimetrică la stânga

În baza acestor relaţii stabilite între medie şi mediană, dar luând în considerare şi abaterea standard, o variantă a coeficientului skewness poate fi

calculată după formula: $Sk = \frac{3(\overline{x}-Me)}{\sigma}$, unde *x barat* reprezintă media, *Me* reprezintă mediana iar σ – abaterea standard. Valorile sale se interpretează la fel ca orice coeficient de asimetrie. Având în vedere simplitatea acestei formule, exersarea ei pe unul dintre exemplele de mai sus, v-o lăsăm ca exercițiu.

Simetria este al doilea aspect important al unei distribuții normale. **O distribuție normală este o distribuție simetrică.**

VI.3 Boltirea sau excesul

Asimetria pe orizontală, presupune, după cum am văzut, o deplasare a tendinței centrale spre stânga sau spre dreapta, către scoruri mici sau către scoruri mari. Aceasta este singura „asimetrie" posibilă? Ei bine, nu. Există şi un fel de „asimetrie verticală" sau boltire. Termenul folosit generic pentru acest concept este termenul de *kurtosis* (din limba greacă, *kurtos* = „cocoşat"). Practic, boltirea se referă la aspectul „cocoaşei" distribuției rezultatelor. Cocoaşa poate fi ascuțită şi atunci putem vorbi de o distribuție ascuțită sau *leptocurtică*, poate fi turtită, distribuția turtită, plată sau *platicurtică* sau normală, distribuție *mezocurtică*. **O distribuție normală este întotdeauna o distribuție mezocurtică.**

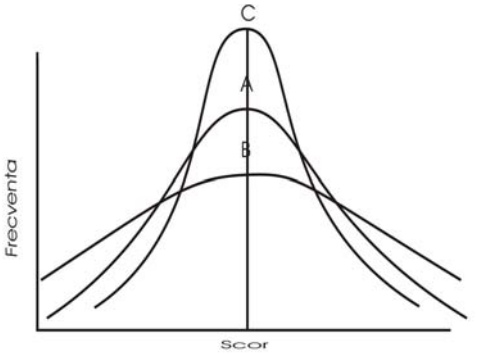


Figura 6.7 – Distribuție mezocurtică (A), leptocurtică (C) şi platicurtică (B)

În figura de mai sus, distribuția „C" este o distribuție leptocurtică, ascuțită. Distribuția „B" este o distribuție platicurtică, turtită iar distribuția „A" este o distribuție normală sub aspectul boltirii, sau mezocurtică.

Aţi observat deja că boltirea nu este altceva decât simetria pe axa verticală (O_Y), spre deosebire de simetria propriu zisă, deplasarea valorilor pe axa orizontală (O_X). Dacă la simetrie vorbeam de frecvenţe (care sunt cele mai frecvente scoruri obţinute, unde se concentrează acestea? În eşantionul nostru avem mai degrabă subiecţi mici sau subiecţi înalţi? Ori poate există un echilibru între subiecţii scunzi şi cei înalţi), la boltire discutăm de valori, de modul în care aceste valori se grupează în jurul tendinţei centrale (sunt mai grupate valorile în jurul mediei sau, din contra, foarte împrăştiate.)

La fel ca şi simetria, boltirea nu poate fi apreciată pur „ochiometric" ci avem nevoie şi aici de anumiţi coeficienţi rezultaţi din formule de calcul specifice. Pearson a discutat despre boltire în termeni de momente, la fel ca şi simetria, rezultând astfel **coeficientul de boltire** sau **coeficientul kurtosis**.

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^4}{n\sigma^4}$$ (formula 6.3)

Dacă skewness reprezenta raportul dintre momentul centrat de ordin trei şi momentul centrat de ordin doi, coeficientul de boltire reprezintă raportul dintre momentul centrat de ordin patru şi momentul centrat de ordin doi.

Formula nu ar trebui să vă pună probleme dacă aţi înţeles modalitatea de calculare a coeficientului de asimetrie. Tehnica este asemănătoare.

Una dintre recomandările interpretative ale coeficientului de boltire, afirmă că, o distribuţie normală are un coeficient de boltire egal cu 3, iar comparaţiile se fac în funcţie de această valoare. Pe măsură ce coeficientul de boltire are valori mai mari de 3, distribuţia este din ce în ce mai cocoşată (leptocurtică), iar cu cât acest coeficient scade de la valori mai mici decât 3, cu atât distribuţia este mai plată (platicurtică).

Deoarece în practica statistică suntem obişnuiţi mai degrabă cu coeficienţi care iau în considerare valori între -1 şi +1, o variantă a formulei de mai sus poate fi următoarea, care-şi păstrează aceeaşi valoare interpretativă:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n\sigma^4} - 3 \qquad \text{(formula 6.4)}$$

Sunt şi alte modalităţi de interpretare ale acestui coeficient, modalităţi pe care le vom discuta imediat, în subcapitolul destinat aplicaţiilor practice.

Vom reconsidera exemplul folosit la aprecierea simetriei, în vederea calculării boltirii. De data aceasta, nu mai suntem interesaţi de ridicarea la cub a diferenţelor dintre valori şi medie, ci de ridicarea la puterea a patra, pentru a afla momentul centrat de ordin patru. După efectuarea calculelor, suma pătratelor diferenţelor este 684,4 iar suma ridicării la puterea a patra a diferenţelor va fi 65316,59. Acum va trebui să aplicăm formula de calcul şi vom obţine valoarea coeficientului kurtosis de 1,39, ceea ce determină o distribuţie uşor platicurtică. Acest lucru devine mai evident dacă aplicam formula a doua şi scădem valoarea 3. Va rezulta un coeficient de boltire de -1,61, putând suspecta această formă a distribuţiei.

Tabel 6.2 – Calculul coeficientului de boltire

x_i	x_i-m	$(x_i\text{-m})^2$	$(x_i\text{-m})^4$
10	-9,6	92,16	8493,466
22	2,40	5,76	33,1776
31	11,40	129,96	16889,6
9	-10,60	112,36	12624,77
24	4,40	19,36	374,8096
27	7,40	54,76	2998,658
29	9,40	88,36	7807,49
9	-10,60	112,36	12624,77
23	3,40	11,56	133,6336
12	-7,60	57,76	3336,218
Suma		**684,4**	**65316,59**

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{65316{,}59}{10 \times \left(\sqrt{\frac{684{,}4}{10}}\right)^4} = \frac{65316{,}59}{46840{,}30} = 1{,}39$$

- O distribuţie **leptocurtică**, ascuţită, arată că datele sunt foarte grupate şi apropiate de medie, lotul de subiecţi având un mare grad de omogenitate a scorurilor. Această distribuţie nu ridică nici un fel de probleme atunci când trebuie să diferenţiem subiecţii care obţin scoruri mici sau mari. Avem însă dificultăţi atunci când trebuie să diferenţiem subiecţii din zona medie a distribuţiei. De exemplu, într-o clasă de 30 de elevi, 27 obţin

medii anuale între 7,9 şi 8,1. Iată că, doar două zecimi diferenţiază între aproape întreg efectivul de elevi. Nu avem nici o problemă cu repartizarea celorlalţi trei elevi. Pe aceia îi vom considera ori foarte buni, ori foarte slabi, în funcţie de media obţinută – sub 7,9 sau peste 8,1. Ce ne facem însă cu cei 27 de subiecţi? Suntem, iată, în imposibilitatea de a-i ierarhiza în vreun fel. Dacă presupunem că cei trei subiecţi au note mai mici de 7,9, atunci cine va lua, dintre cei 27 de elevi, premiul I, cine va lua premiul II şi cine va lua premiul III. Decizia ar fi, după cum se poate vedea, extrem de dificilă, dacă nu chiar imposibilă. Singura variantă în acest caz, ar putea fi creşterea preciziei. Nu calculăm media cu o singură zecimală, ci cu 2, 3 sau 4 zecimale. Totuşi, sunt situaţii în care un asemenea nivel de precizie este ridicol. O evaluare în care distribuţia rezultatelor este leptocurtică, este o evaluare ce nu poate diferenţia între subiecţii de nivel mediu, iar o curbă leptocurtică nu este o distribuţie normală;

- O distribuţie **platicurtică**, plată, este o distribuţie în care rezultatele sunt foarte împrăştiate faţă de medie şi indică un grad ridicat de eterogenitate al scorurilor. Problema generală a acestei distribuţii, în opoziţie cu distribuţia leptocurtică, este aceea că diferenţiază greu la extreme şi destul de bine în zona mediei. Va fi greu, utilizând o asemenea distribuţie, să facem diferenţieri între elevii slabi şi între elevii buni, deşi putem diferenţia relativ uşor elevii medii. Datorită acestui fapt, o distribuţie platicurtică nu este nici ea o distribuţie normală.

- **O distribuţie normală este o distribuţie mezocurtică.**

Analiza unei distribuţii sub aspectul normalităţii este primul pas pe care-l facem în orice prelucrare de date. Deoarece, în funcţie de rezultatul

acestei analize, vom putea alege tehnicile şi procedeele statistice pe care le putem folosi, această etapă o întâlnim, de obicei, la începutul oricărui raport de cercetare, imediat după descrierea eşantionului.

Deşi modalitatea, simetria şi boltirea se pot analiza separat, în practică putem avea surpriza ca cei doi indicatori ai simetriei şi boltirii să-şi cumuleze efectele şi, cu toate că separat putem aprecia o distribuție ca fiind normală, luată per ansamblu vom constata că distribuția nu este normală. Există, aşadar, alte tehnici de verificare a distribuției normale, tehnici pe care le vom discuta în ultimul capitol al acestui volum deoarece mai aveți nevoie de unele informații suplimentare.

VI.4 Caracterizarea distribuției normale

Afirmam, în capitolul anterior, că media şi abaterea standard caracterizează pe deplin o distribuție normală. Teoretic, există un număr infinit de abateri standard la dreapta şi la stânga mediei. Practic, considerăm că o distribuție normală are un număr de şase abateri standard: trei la dreapta mediei şi trei la stânga mediei, deoarece volumul de populație care rămâne în afara acestor limite este nesemnificativ şi poate fi ignorat. La o distribuție perfect normală, jumătate din rezultate se situează în dreapta mediei (rezultatele mari) şi jumătate din rezultate se situează în stânga mediei (rezultate mici).

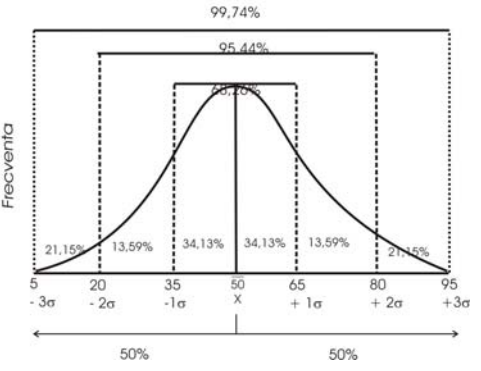


Figura 6.8 – Distribuția normală

- Între o abatere standard la stânga mediei şi o abatere standard la dreapta mediei, întâlnim aproximativ 68% din populație (mai precis 68,28%, cu 34,13% între medie şi o abatere standard la stânga şi tot 34,13% între medie şi o abatere standard

la dreapta). Aceasta este **zona normală**, zona în care rezultatele pot fi considerate tipice. Dacă luăm, ca exemplu, coeficientul de inteligență, în această zonă se află persoanele considerate normale sub aspect intelectual;

- Între două abateri standard la stânga mediei şi două abateri standard la dreapta, întâlnim aproape 95% din populație (mai precis 95,44% cu 47,72% între medie şi două abateri standard la stânga şi 47,72% între medie şi două abateri standard la dreapta.). Zona între o abatere standard şi două abateri standard, atât la stânga cât şi la dreapta, este **zona rezultatelor accentuate** şi cuprinde, fiecare, câte 13,59% din populație. Astfel, între o abatere standard la stânga şi două abateri standard la stânga, de află persoanele cu un intelect mai redus, intelect de limită. La fel, între o abatere standard la dreapta şi două abateri standard la dreapta, se află persoanele cu o inteligență peste medie, cei „deştepți";
- Între trei abateri standard la stânga mediei şi trei abateri standard la dreapta mediei, întâlnim aproximativ 99% din populație (mai precis 99,74% cu 49,87 % între medie şi trei abateri standard la stânga şi 49,87% între medie şi trei abateri standard la dreapta). Această zonă este zona **rezultatelor atipice.** Zona între două abateri standard la stânga şi trei abateri standard la stânga, este zona deficienților mintali, cu un intelect extrem de redus. În această zonă, întâlnim 2,15% din populație, acelaşi procent ca şi în zona situată între două şi trei abateri standard la dreapta, zona celor cu o inteligență deosebită, a persoanelor supradotate. Iată că, între deficienții mintali şi supradotați întâlnim aproape întreaga populație, peste 99%

- Mai rămâne, aproximativ 1% din populație (mai exact 0,26%). situată fie după trei abateri standard la stânga (0,13%), fie după trei abateri standard la dreapta (0,13%). **Iată şi zona rezultatelor aberante**, zonă în care găsim mai puțin de 1% din populație. După trei abateri standard la stânga, găsim… idiotul idioților, iar după trei abateri standard la dreapta îl găsim pe Einstein. Iată, această zonă este zona idioției sau a geniilor.

Spuneam că, dacă noi cunoaştem media şi abaterea standard, putem calcula probabilitatea cu care un rezultat particular se încadrează într-una dintre aceste zone. Să considerăm că, figura de mai sus, reprezintă o distribuție perfect normală a rezultatelor unui test, la care scorurile variază de la 0 la 100. Un subiect care obține nota brută 64, este situat în zona medie, foarte aproape de rezultatele zonei superioare, în timp ce un subiect cu nota brută 27 se situează în zona accentuată inferioară. Iată că, astfel, ajungem la un alt concept în statistică şi anume la conceptul poziție a unui scor particular în raport cu rezultatele unui grup reprezentativ de persoane (eşantion). Oare nu cumva acest lucru seamănă foarte bine cu un alt termen vehiculat în psihologie şi anume termenul de etalon? Ba da, ați presupus foarte bine, chiar despre etaloane este vorba. Însă, până să le aprofundăm, vom aborda un alt element de bază şi anume conceptul de note „z".

VI.4.1 Notele „z" şi note standard derivate

Când am discutat de abaterea standard, am văzut că aceasta reprezintă un mijloc de raportare a modului în care rezultatele se „împrăştie" în jurul mediei. De asemenea, am spus că, dacă cunoaştem media şi abaterea standard a unei distribuții, putem foarte uşor să precizăm probabilitatea ca un scor particular să se situeze într-o anumită poziție. Pentru disciplinele socio-umane, acest lucru este vital, deoarece atunci când măsurăm, practic precizăm poziția pe care o are un subiect în raport cu o populație de referință. Este ideal ca această poziție să poată fi prezentată într-o manieră standardizată. Acest lucru

poate fi uşor de îndeplinit atunci când avem în vedere conceptul de note standardizate. La baza acestui concept stau notele „z". **Notele „z" reprezintă diferența dintre scorul observat şi medie, în termeni de abatere standard.** Cu alte cuvinte, notele „z" nu sunt altceva decât distanțele la care se situează scorurile particulare în raport cu media grupului iar această distanță este exprimată standardizat.

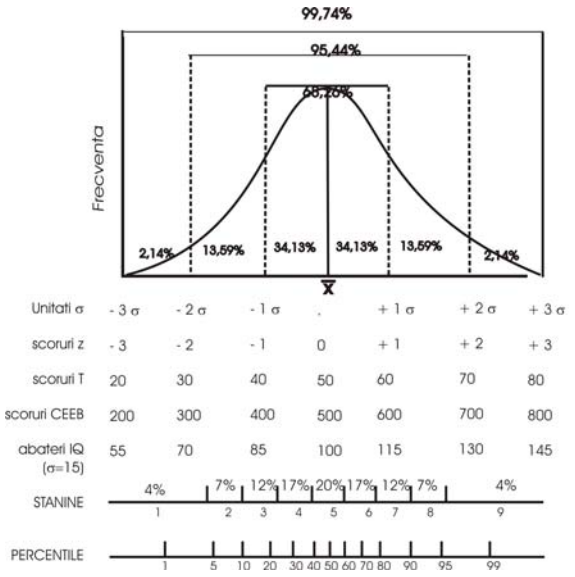


Figura 6.9 – Diferite sisteme de note standard

Teoretic, notele „z" sunt note obținute pe o „curbă" cu media 0 şi abaterea standard 1. În acest caz, o distribuție normală are practic notele „z" cuprinse între -3 şi +3, între aceste note regăsindu-se peste 99% din populație, după cum observați în figura de alături.

Transformarea notelor obținute de către un subiect în note „z" este foarte simplă şi se face după formula:

$$z = \frac{x - \overline{x}}{\sigma} \quad \text{(formula 6.5)}$$

în care *x* reprezintă nota brută, scorul obținut de către subiect, *x barat* reprezintă media eşantionului cu care comparăm performanța subiectului iar *sigma* se referă la abaterea standard a acelui eşantion.

De exemplu, într-o şcoală, media elevilor unei clase la matematică este de 6,25 cu o abatere standard de 3,94. Georgel a obținut, la matematică, media 7,15. Ne interesează să ştim dacă Georgel este un elev mediu la matematică, un elev bun sau un elev slab. Intuitiv, am putea spune că este un elev bun, deoarece media lui este mai mare decât media clasei. Nu putem şti însă cu precizie acest lucru, deoarece nu cunoaştem cum se distribuie notele celorlalți elevi în jurul mediei. Acest lucru ni-l spune abaterea standard. Cu alte cuvinte, va trebui să precizăm poziția pe care o ocupă Georgel în cadrul grupului de referință din care face parte, adică raportat la elevii din clasa sa. Acest lucru devine posibil prin calculul notei „z" a lui Georgel, notă care ne spune la câte abateri standard se află scorul său în raport cu clasa. Aplicând formula de mai sus, observăm că nota „z" a lui Georgel la matematică este de $\frac{7,15-6,25}{3,94} = \frac{0,9}{3,94} = 0,22$. Deci Georgel se situează, din punctul de vedere al performanței sale la matematică, la 0,22 abateri standard față de medie, adică în zona punctului roşu marcat pe figura 6.10.

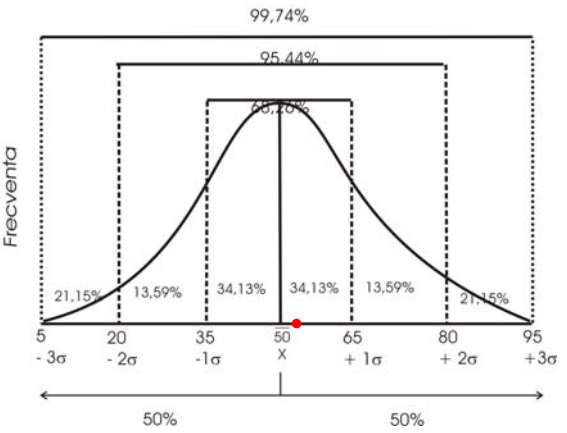


Figura 6.10 – Reprezentarea poziției unui subiect în funcție de nota z

Performanța sa la matematică este, prin urmare, o performanță medie, situându-se între medie şi o abatere standard la dreapta, deci în zona scorurilor medii.

Spre deosebire de Georgel, Costel are media la matematică de 4,08. Este oare Costel un elev slab, un elev mediu sau un elev foarte slab? Nota „z" a lui Costel este – 0,55. Şi această notă se situează între medie şi o abatere standard la stânga de această dată. Iată că, în ciuda aparențelor, raportat la

performanța elevilor din clasă, deşi are media sub cinci, Costel este un elev mediu la această disciplină.

Cred că intuiți deja ce se întâmplă. În domeniul socio-uman, noi nu putem niciodată să măsurăm ceva cu exactitate. Putem doar să comparăm rezultatul unui subiect cu rezultatul unui grup de referință. Comparația aceasta se poate realiza numai dacă cunoaştem media şi abaterea standard a grupului de referință, în cazul în care, distribuția caracteristicii măsurate în cadrul grupului de referință este o distribuție normală.

Pentru a vă clarifica mai bine acest concept, să luăm un nou exemplu. Avem un test de inteligență cu 20 de probleme. Dacă subiectul rezolvă corect o problemă, primeşte un punct. Dacă nu o rezolvă corect, primeşte zero puncte. Deci, scorul unui subiect poate varia de la zero puncte, atunci când nu rezolvă corect nici o problemă, şi până la 20 de puncte atunci când le rezolvă corect pe toate.

Avem două persoane: Vasile, obține un scor la acest test de 10 puncte iar Viorel, un scor de 12 puncte. Întrebare: puteți spune care dintre cei doi subiecți este mai bun? Care are un nivel intelectual mai ridicat?

Cei mai mulți dintre dumneavoastră vor zâmbi acum complice şi îmi vor spune: „Evident că Viorel, deoarece a rezolvat mai multe probleme". Poate aveți dreptate. Însă, iată că vă mai dau un amănunt. Vasile este cioban iar ciobanii rezolvă corect, în medie, 7 probleme. Lucrurile se schimbă acum. Vasile poate obține liniştit funcția de baci deoarece nivelul său intelectual este mai ridicat în comparație cu nivelul ciobanilor. Viorel este profesor iar profesorii rezolvă corect, în medie, 18 probleme. Față de grupul profesorilor, Viorel este… cioban.

Deşi exagerat, scopul acestui exemplu este acela de a vă convinge că notele brute, scorurile, nu ne spun nimic prin ele însele. Notele brute dobândesc semnificație numai în urma comparării cu o populație de referință. Pu-

tem afirma despre cineva că este bun sau slab numai în raport cu un reper, cu un grup de referință. Probabil că eu sunt bun la statistică în comparație cu studenții cărora le predau această disciplină şi slab în raport cu unii profesori de ai mei sau cu Spearman ori Pearson. Termenul de bun sau slab dobândeşte semnificație numai raportat la un reper.

Toate aceste informații servesc drept suport procesului de etalonare. Un etalon nu este altceva decât un sistem de norme, la care se raportează performanța unui individ, la o anumită sarcină. Evident, un etalon este făcut pe o anumită populație iar raportând un rezultat al unui subiect la un etalon, nu facem altceva decât să comparăm performanțele acelui subiect cu performanțele grupului pe care a fost construit etalonul. Despre etaloane, vom discuta însă imediat.

Notele „z" au şi ele o serie de proprietăți cu aplicabilitate practică deosebită, dintre care menționăm (Popa, 2006):

- Media unei distribuții „z" este egală cu zero, afirmație ce rezultă din proprietatea acestui indicator de a se diminua atunci când scădem o constantă din fiecare scor particular. Deoarece din formulă rezultă această diferență (se scade fiecare scor particular din medie), în final media va ajunge la valoarea zero
- Abaterea standard a unei distribuții „z" este întotdeauna 1. Şi această afirmație rezultă din proprietățile abaterii standard. Ştim că dacă împărțim abaterea standard la o constantă, valoarea acesteia se divide corespunzător. Din formulă, observăm că acea constantă la care împărțim este chiar abaterea standard iar împărțirea a două numere egale are ca rezultat 1

Notele „z" sunt note direct calculabile, utilizându-se media şi abaterea standard şi reprezintă „temelia" oricărui proces de standardizare. Totuşi, principalul dezavantaj al notelor „z" este acela că sunt puțin intuitive. Trebuie

să ştim bine ce înseamnă distribuția normală pentru a înțelege corect semnificația acestor note. În plus, au valori pozitive şi negative, ceea ce le face, iarăşi, greu de utilizat.

Ar fi mai uşor pentru noi dacă în loc să folosim o scală de la – 3 la 3 pentru încadra aproximativ 99% din populație, am folosi o altă scală, de exemplu de la 1 la 10, mult mai intuitivă, deoarece se asociază rapid sistemului şcolar de evaluare sau o scală de la 1 la 5, de la 1 la 9 şi aşa mai departe. Este oare posibil acest lucru? Dacă ne amintim de proprietățile scalei de interval, prezentate la începutul acestei lucrări, vom constata că una dintre aceste proprietăți se referea la deplasare, prin modificarea constantei de deplasare. Notele „z" au o constantă de deplasare zero. Dacă mărim această constantă de deplasare, vom putea să „deplasăm" distribuția astfel încât să nu ne mai situăm în zona scorurilor standardizate negative. Obținem astfel un alt sistem de note standardizate, mai intuitiv şi mai practic. Asemenea scale standardizate, derivate din notele „z", sunt **staninele** (numele provine din englezescul standard nine), un sistem în nouă clase normalizate, **sten** (din englezescul standard ten), în zece clase normalizate, **notele „T"** – sistem raportat la o distribuție cu media 50 şi abaterea standard 10, **abateri IQ** – media 100 şi abaterea standard 15, etc.

Reluând exemplele anterioare, ne propunem să convertim scorurile „z" ale lui Georgel şi Costel în stanine, sten, note T şi abateri IQ. Vă reamintesc că media elevilor clasei la matematică este de 6,25 cu o abatere standard de 3,94 iar Georgel a obținut, la matematică, media 7,15 şi Costel 4,08.

Pentru a calcula scorurile sten, vom deplasa distribuția cu media scalei sten care este, evident, 5,5 iar formula de calcul va deveni:

$$sten = 5{,}5 + \frac{x-\overline{x}}{\sigma} = 5{,}5 + z. \quad \text{(formula 6.6)}$$

Aşadar, scorul sten al lui Georgel este de 5,5+0,22, adică 5,72. Observăm că poziția acestui scor nu modifică punctul în care se află Georgel pe

distribuția normală, însă scorul devine mult mai intuitiv. Media 5,75 la matematică ne spune mult mai multe decât 0,22. Scorul în stanine al lui Georgel se calculează similar. Pe o scală de la 1 la 9, media este evident 5. Prin urmare, deplasăm scala cu 5 de această dată şi obținem 5 + 0,22 = 5,22. Obținem, de fapt, exact acelaşi lucru, însă pe o scală de la 1 la 9 şi nu de la 1 la 10 ca în exemplul anterior.

Pentru notele „T", avem de a face atât cu o deplasare, cât şi cu o comprimare a intervalelor. Afirmam că scala „T" este o scală cu media 50 şi abaterea standard 10 iar în acest caz formula de calcul a notelor „T" va deveni:

$$T = 50 + 10\frac{x-\overline{x}}{\sigma} = 50 + 10z. \text{ (formula 6.7)}$$

Nota „T" a lui Georgel, pe o scală de la 1 la 50 va fi aşadar 50 + 10x0,22, adică 50 + 2,2 ceea ce înseamnă 52,2. Surpriză. Din nou, nu se modifică poziția acestui scor pe curba de distribuție. La fel putem proceda şi în cazul calculului abaterilor IQ. Formula, de această dată va fi similară celei pentru note „T":

$$Ab_{IQ} = 100 + 15\frac{x-\overline{x}}{\sigma} = 100 + 15z. \text{ (formula 6.8)}$$

În cazul nostru, rezultatul devine 100 + 15x0,22 adică 100 + 3,3 ceea ce înseamnă 103,3. Remarcăm că, nici în acest caz, nu se modifică poziția subiectului pe curba de distribuție.

Iată cât de uşor putem raporta performanțele subiectului la performanțele grupului din care acesta face parte şi care sunt semnificațiile acestei raportări, indiferent de sistemul de scale standard utilizat. Calculul notelor standardizate derivate ale lui Costel vi-l lăsăm dumneavoastră ca exercițiu.

O altă caracteristică extrem de importantă a scorurilor standardizate este aceea că, putem efectua comparații între rezultatele obținute de către subiecți la probe diferite care folosesc ele însele scale diferite. De exemplu,

dacă comparăm performanţele unui subiect la un test de inteligenţă, care are 60 de probleme cu performanţele obţinute de acelaşi subiect la o probă de comprehensiune semantică, ce are 30 de probleme iar subiectul obţine o notă brută la testul de inteligenţă de 40 iar la cel de comprehensiune semantică de 28 atunci putem oare afirma că subiectul are o inteligenţă mai bună în comparaţie cu capacitatea sa de comprehensiune semantică? Categoric nu! Chiar şi la prima vedere, în condiţiile în care ambele teste au acelaşi sistem de scorare, constatăm că primul are mai multe întrebări decât al doilea. Sunt 20 de probleme nerezolvate de subiect la testul de inteligenţă în comparaţie cu 2 nerezolvate la cel de comprehensiune semantică. Este imposibil să comparăm direct aceste două rezultate, mai întâi trebuie să le aducem la un numitor comun. Acest numitor comun îl reprezintă notele standardizate.

Prin urmare, va trebui să calculăm notele „z" (sau orice alte note standardizate) ale acestui subiect la cele două teste şi apoi vom putea să comparăm rezultatele obţinute utilizându-se aceste note standardizate. De nenumărate ori mi-a fost dat să văd lucrări ale studenţilor prin care aceştia comparau rezultatele subiecţilor la două teste diferite atât sub aspectul numărului de itemi cât şi sub cel al sistemului de cotare. E ca şi cum aş compara două kilograme de mere cu 2 metri de sfoară. O asemenea abordare este complet eronată şi nu duce la nici un fel de rezultat. Trebuie găsit un numitor comun pe care îl regăsim sub forma acestor note standardizate.

VI.5 Transformarea datelor brute

Consider că v-am convins de avantajele utilizării notelor standardizate. Totuşi, folosirea lor este condiţionată de o distribuţie normală sau foarte apropriată de cea normală. În absenţa criteriului normalităţii, tot demersul menţionat mai sus nu se mai susţine. Am văzut că sunt destule situaţii în care distribuţia rezultatelor nu mai este normală, putând apărea scoruri aberante ce duc la o asimetrie pronunţată sau un eşantion extrem de omogen sau extrem de eterogen ce poate influenţa radical boltirea. De asemenea, prezenţa mai

multor „vârfuri" în distribuție determină încălcarea criteriului modalității. Ce ne facem în aceste cazuri? Vestea bună este că avem la dispoziție metode prin care putem normaliza această distribuție făcând-o utilizabilă.

Am văzut deja că, prin inspectarea datelor, putem vedea şi corecta scorurile extreme rezultate din erorile de introducere a datelor. Dacă aparatele de înregistrare sunt defecte ori designul de cercetare este defectuos, nu avem prea multe posibilități decât reproiectarea cercetării sau repararea aparatelor. Totuşi, ce facem dacă nu sunt erori de introducere a datelor, aparatele sunt bune iar designul este în regulă, însă nu am obținut o distribuție normală?

Soluția este transformarea datelor brute prin anumite procedee de rescalare a valorilor, pentru ca noile valori să respecte condițiile de aplicare a unor teste statistice (Sava, 2004). Dacă şi în urma acestor transformări distribuția continuă să pună probleme, vom utiliza statistici nonparametrice, renunțând la prelucrările parametrice de date.

Există multe procedee de modificare a valorilor care să nu afecteze semnificația datelor, iar dintre cele mai cunoscute sunt tehnicile de normalizare a distribuției.

O distribuție multimodală poate fi corectată relativ uşor inspectând valorile modale şi renunțând la un singur scor. Deoarece modul este categoria cu frecvența cea mai mare, renunțând la un singur caz din categoria modală transformăm o distribuție bimodală într-una unimodală. De exemplu dacă măsurăm greutatea unui număr de 50 de subiecți şi constatăm că avem două categorii mod: subiecți cu greutatea de 78 de kilograme – 10 cazuri şi subiecți cu greutatea de 84 de kilograme – 10 cazuri, eliminând din eşantion un singur subiect care are 78 de kilograme sau 84 de kilograme, determinăm transformarea într-o distribuție unimodală. Desigur, nu vom mai avea 50 de cazuri ci 49, însă este mai bine să tai un deget decât să pierzi mâna.

O distribuţie asimetrică sau excesivă presupune utilizarea unor tehnici diferite, dintre care menţionăm:

- **Folosirea mediei 5% trim** – este o tehnică ce utilizează o medie ajustată în care nu se ţine cont de 5% dintre scorurile aflate la extremităţile distribuţiei. Metoda se foloseşte în cazul în care avem scoruri extreme reale. Prin acest procedeu, ele sunt eliminate din calculul acestui indicator;
- **Extragerea radicalului din toate valorile distribuţiei** este o tehnică folosită în special atunci când avem o distribuţie asimetrică moderată. Radicalul „strânge" o distribuţie puternic platicurtică şi corectează asimetria;
- **Ridicarea la pătrat sau la cub** este o tehnică utilizată în cazul distribuţiilor leptocurtice deoarece prin acest procedeu se „împrăştie" rezultatele în jurul tendinţei centrale;
- **Logaritmarea valorilor individuale şi inversarea valorilor** sunt alte procedee de normalizare utilizate în cazul unor distribuţii asimetrice severe.

După aceste prelucrări de date este necesară, în mod obligatoriu, o reverificare a distribuţiei pentru a vedea modul în care se respectă criteriile normalităţii distribuţiei pe date astfel normalizate. Toate aceste operaţii sunt permise fără a afecta semnificaţia iniţială a datelor, dacă luăm în considerare proprietăţile scalei de interval discutate în primul capitol.

Pentru a clarifica aceste situaţii, vom considera o serie de exemple intuitive propuse de F. Sava (Sava, 2004) şi pe care am apreciat că ar fi util să le reproducem aici.

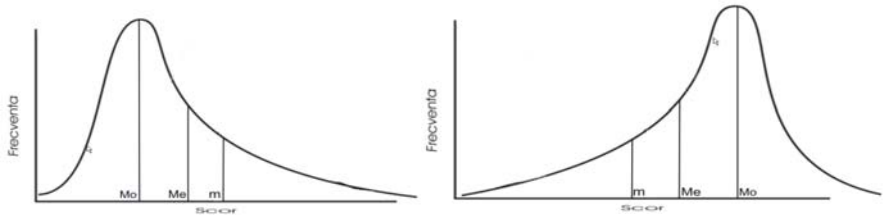


Figura 6.11 – Distribuţie asimetrică moderată

Cele două figuri de mai sus indică o asimetrie pozitivă, respectiv negativă moderată. În primul caz, extragerea rădăcinii pătrate din fiecare scor individual va conduce la normalizarea acestei distribuţii. În al doilea caz radicalul se extrage din reflexia scorurilor. Prin operaţia de reflectare, transformăm o asimetrie negativă într-o asimetrie pozitivă, după care putem extrage rădăcina pătrată.

Operaţia de reflectare este una extrem de uşoară şi presupune practic inversarea valorilor. Dacă, de exemplu, avem următorul şir de date:

Şirul normal: 5, 9, 13, 8, 7, 19

Primul pas pe care trebuie să îl facem este acela de a identifica valoarea maximă din şirul de date. În cazul nostru, valoarea cea mai mare din şir este 19. Adăugăm apoi 1 la valoarea maximă şi obţinem valoarea 20.

La al doilea pas, vom scădea din valoarea obţinută (20) fiecare scor şi vom obţine şirul reflectat. Constatăm că această operaţiune este similară unei reflexii în oglindă, în care stânga devine dreapta şi dreapta stânga. Procedura este necesară în cazul distribuţiilor asimetric negative în vederea extragerii radicalului, în vederea logaritmării sau pentru inversare.

Şirul reflectat: 15, 11, 7, 12, 13, 1

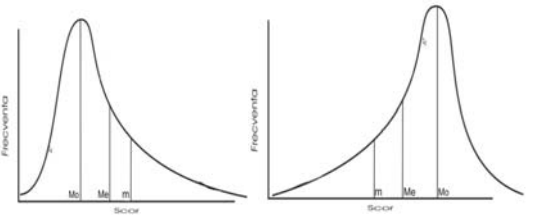


Figura 6.12 – Distribuţie asimetrică severă

O asimetrie ca cea din figura de mai sus reprezintă o asimetrie severă. În acest caz, extragerea radicalului din fiecare valoare sau din valorile reflectate nu rezolvă satisfăcător problema. Avem nevoie, similar situaţiei de mai sus, de logaritmul natural din fiecare valoare sau, în cazul asimetriei negative, de logaritmul natural din fiecare valoare inversată.

În cazul asimetriilor extreme curba are un aspect de curbă în „J", cu asimetrie pozitivă sau negativă. În această situaţie nici logaritmul natural nu ne mai rezolvă problema. Vom proceda la calcularea inversului fiecărei valori pentru a normaliza distribuţia. Inversul unei valori îl putem obţine împărţind 1 la fiecare valoare din şirul de date. Cu alte cuvinte, inversul valorii X este dat de $^{1}/_{X}$. Această inversare se realizează folosind valorile din şirul nostru de date dacă asimetria este pozitivă sau valorile din şirul inversat de date dacă asimetria este negativă.

VI.6 Sisteme de etalonare

Mulţi psihologi consideră că realizarea unui etalon înseamnă, de fapt, adaptarea unei probe psihologice la specificul unei populaţii. Nimic mai fals. Etalonarea reprezintă ultima etapă în procesul de adaptare a unui instrument sau de construcţie a unei probe noi şi nu are nici o treabă cu celelalte calităţi psihometrice ale unui test. **Un etalon reprezintă un sistem de norme prin care putem compara rezultatele unui subiect (caz individual) cu rezultatele obţinute de către un grup reprezentativ de subiecţi.** Un etalon este, aşadar, o unitate de măsură, care se foloseşte în acelaşi mod în care utilizăm

orice aparat de măsurare. Atunci când măsurăm greutatea unei persoane, avem valoarea acestei variabile obținută prin folosirea unui cântar. Însă semnificația acestei valori depinde de sistemul de referință la care ne raportăm. Dacă o persoană are 100 de kilograme şi se raportează la obezi, cu o greutate peste 120 de kilograme, atunci acea persoană va fi una slabă. Dacă însă acelaşi individ îl comparăm cu persoane de 70-80 de kilograme, o putem considera o persoană „grea".

Mai delicate sunt însă lucrurile în ştiințele socio-umane. Dacă atunci când ne referim la greutate avem totuşi un zero absolut iar un kilogram înseamnă un kilogram indiferent despre ce am vorbi, în ştiințele sociale lucrurile nu stau chiar aşa. Ştiți deja de ce, astfel încât nu vom mai insista asupra acestor aspecte. În domeniul nostru, măsurarea înseamnă de fapt ierarhizare. Secțiunea anterioară a lămurit modul în care comparăm performanța unui subiect cu performanța unui grup. Nu avem cum să facem în permanență acest lucru. Dacă lucrăm într-o întreprindere şi trebuie să evaluăm un strungar nou angajat, atunci, pentru evaluarea acestuia, nu va trebui să reexaminăm toți strungarii din întreprindere pentru ca să comparăm performanța acestuia cu performanța grupului din care face parte. Ar fi absurd. Vom considera că performanța grupului rămâne stabilă un anumit interval de timp. În aceste condiții, examinând la un moment dat grupul de referință, putem reține concluziile desprinse şi apoi putem compara orice nou individ cu aceste concluzii. Ne referim acum tocmai la conceptul de etalon.

Etaloanele se construiesc în funcție de aspectul distribuției datelor populației pe care s-a făcut studiul. Pentru distribuții care **se abat de la normalitate**, singura metodă de realizare a unui etalon este cuantilarea, utilizându-se valorile percentile. Deoarece se bazează pe poziția valorilor în şirul ordonat de date, similar medianei şi cuartilelor, asemenea etaloane sunt considerate ca fiind etaloane „slabe" şi au o serie de limitări pe care le vom discuta. **O distribuție normală**, în care media este un indicator bun al tendinței centrale, permite construcția unor etaloane normalizate, în unități sigma sau

folosindu-se notele „z”, mult mai precise şi mai utile în comparaţie cu etaloanele bazate pe percentile.

Înainte de a intra în detalii tehnice referitoare la construcţia unui etalon, să menţionăm câteva criterii ce indică un etalon bun, precum şi aspectele pe care le urmărim atunci când într-o probă psihologică ni se oferă un asemenea sistem de norme.

- Un etalon bun este un etalon construit pe **un număr suficient de mare de persoane**. Deoarece scopul unui etalon este acela de a compara un caz individual cu o populaţie de referinţă, această populaţie de referinţă trebuie să includă un număr suficient de mare de subiecţi astfel încât să fie reprezentativă. Cunoaştem deja problemele pe care le poate pune un eşantion redus ca dimensiuni. Din nefericire acestea nu sunt singurele. Există şi alte aspecte pe care le vom discuta ulterior. Un număr mic de persoane selectate în eşantionul pe baza căruia se construieşte etalonul – numit şi *eşantion normativ* – duce la o reprezentativitate scăzută a acestuia. Comparând apoi un subiect cu acest eşantion, avem toate şansele să ajungem la concluzii complet eronate. Un eşantion prea mare creşte considerabil costul etalonării făcând cercetarea extrem de scumpă. Trebuie astfel găsit un optim între numărul subiecţilor incluşi în eşantionul normativ şi reprezentativitatea acestuia. Un etalon bun este un etalon construit pe un număr de 250-300 de subiecţi. Se acceptă, în anumite situaţii şi etaloane construite pe mai puţin subiecţi dacă acestea sunt specifice unui anumit grup (de vârstă, profesional, clinic etc.). Etaloanele testelor profesionale se construiesc pe mii de subiecţi. De exemplu, eşantionul normativ pe care a fost etalonat chestionarul 16PF de către Pitariu şi Iliescu a cuprins cca. 3000 de subiecţi de pe întreg teritoriul României;

- Un etalon bun este un **etalon adaptat specificului probei psihologice şi a condiţiilor de evaluare**. Un etalon pe populaţie generală are o valoare limitată într-un scop specific, spre exemplu într-o uzină. De aceea se recomandă efectuarea de etaloane specifice atunci când situaţia o impune;
- Un etalon bun are un **număr suficient de mare de clase** care să poată diferenţia subiecţii. Un etalon cu două clase, nu ne este prea util. O împărţire a subiecţilor în „buni" şi „slabi" este rigidă şi fără o mare valoare informativă. Ceva mai bine ar sta lucrurile dacă am avea trei clase: „buni", „medii" şi „slabi", doar că şi în acest caz vom avea o încadrare oarecum forţată a subiecţilor în cele trei clase. Putem discuta de un etalon cu o capacitate acceptabilă de a diferenţia subiecţii de la cinci clase în sus. La cinci clase vom avea următoarele niveluri: „foarte slab", „slab", „mediu", „bun", „foarte bun", niveluri ce permit ordonarea mai fină a subiecţilor. Un număr prea mare de clase însă poate crea confuzii iar o asemenea precizie devine, din nou, exagerată. În psihologie se utilizează, în general, etaloanele în cinci clase, în şapte clase, în nouă clase sau în unsprezece clase.

De multe ori, atunci când un psiholog primeşte şi utilizează o probă psihologică, nu îşi pune probleme asupra etalonului furnizat şi îl acceptă ca atare. Mai mult, unele etaloane sunt pur şi simplu scrise de mână pe o bucată de hârtie, circulă între psihologi, iar aceştia le acceptă necondiţionat. Nimic mai eronat. Nu uitaţi că etalonul este „unitatea noastră de măsură". Avem obligaţia să obţinem informaţii legate de construcţia etaloanelor, cine şi cum le-a făcut, deoarece numai în acest fel vom putea avea încredere în instrument şi vom trage concluzii diagnostice pertinente. De fiecare dată când primiţi un etalon, acesta trebuie să conţină, în mod obligatoriu, următoarele informaţii:

- **Structura eşantionului pe care s-a realizat etalonul**, sub aspectul variabilelor de eşantionare. Etalonul este realizat pe populaţie generală sau pe subiecţi care fac parte dintr-o anumită categorie? Un etalon realizat pe studenţi sau pe elevi de liceu va avea o valoare mai mult decât discutabilă într-o uzină chimică. Care este vârsta şi genul biologic al subiecţilor? Care este distribuţia acestor variabile în structura eşantionului? Dacă avem un etalon realizat pe copii între 10 şi 14 ani, cum oare îl vom putea folosi în studiile asupra bătrânilor? Nu există o regulă asupra variabilelor de eşantionare utilizate la construcţia unui etalon. În general se foloseşte vârsta, sexul, nivelul de şcolarizare. Indiferent de variabilele folosite de către realizatorul etalonului, acesta trebuie să le prezinte. Nu ne putem mulţumi doar cu sintagma „etalonul a fost realizat pe populaţie generală". Trebuie precizată şi structura acestei populaţii. Numărul subiecţilor din eşantion va trebui de asemenea bine precizat. Un etalon realizat pe 100 de subiecţi înseamnă cu totul altceva în comparaţie cu un etalon realizat pe 1000 de subiecţi.
- **Distribuţia răspunsurilor subiecţilor la funcţia investigată** are o deosebită importanţă. Ne interesează să ştim dacă distribuţia a fost normală sau dacă s-au aplicat metode de normalizare a distribuţiei, caz în care trebuie prezentate ambele distribuţii. Suntem interesaţi să aflăm dacă, de exemplu, în eşantion majoritatea persoanelor au rezolvat corect 18 sarcini din 20 sau, din contra, au realizat corect 6 sarcini din 20. Acest lucru ne informează asupra nivelului populaţiei cu care vom compara rezultatele cazurilor individuale;

- **Metoda de etalonare folosită** se stabileşte în funcţie de distribuţia rezultatelor. La o distribuţie normală vor fi folosite, în general, unităţile sigma sau clasele normalizate. În cazul unei distribuţii care se abate de la normalitate, se utilizează sistemul de etaloane bazat pe valorile percentile;
- **Data la care a fost realizat etalonul** ne furnizează informaţii despre momentul în care a fost construit. Datorită dinamicii sociale şi a noilor achiziţii tehnologice, atitudinile, normele sociale se modifică rapid, astfel încât un etalon devine depăşit într-un timp destul de scurt. Un adolescent de astăzi, examinat, de exemplu, cu Inventarul Multifazic de Personalitate Minnesota, ar intra cu uşurinţă în normele patologice din anii 1950. Un etalon, în general, se reactualizează după 5-6 ani.

VI.6.1 Etaloane în cuantile

Au la bază valorile percentile şi se stabilesc într-un mod similar calculului medianei sau al cuartilelor. Bazându-se pe poziţia valorilor în cadrul unui şir ordonat de date, aceste etaloane nu au precizia celor în clase normalizate însă au avantajul că pot fi construite foarte uşor şi pe distribuţii care se abat de la distribuţia normală.

Construcţia unor asemenea etaloane se realizează după următorul algoritm:

- Se stabileşte minimum şi maximum dintr-o distribuţie;
- Se înscriu toate valorile situate între minimum şi maximum;
- Se stabilesc frecvenţele absolute şi cumulate;
- Se alege numărul de clase al etalonului;
- Se calculează scorul corespunzător procentului pentru fiecare limită de clasă.

Să luăm, spre exemplu, următorul şir de date, care reprezintă înălțimea unor subiecți exprimată în centimetri: *182, 175, 174, 189, 177, 177, 180, 173, 188, 171, 184, 192, 180, 172, 177, 193, 184, 192, 172, 177, 174, 171, 172, 179, 185, 179, 173, 172, 179, 188*

În acest exemplu, minimul este 171 iar maximul este 193. Acum, vom ordona crescător toate valorile situate între minim şi maxim şi vom stabili frecvențele absolute şi cumulate.

Tabel 6.3 – Realizarea unui etalon în cuantile

Valoare	**Diagrama în linii**	**f**	**f_c**
171	||	**2**	**2**
172	||||	**4**	**6**
173	|	**1**	**7**
174	||	**2**	**9**
175	|	**1**	**10**
176		**0**	**10**
177	||||	**4**	**14**
178		**0**	**14**
179	|||	**3**	**17**
180	||	**2**	**19**
181		**0**	**19**
182	|	**1**	**20**
183	|	**1**	**21**
184	||	**2**	**23**
185	|	**1**	**24**
186		**0**	**24**
187		**0**	**24**
188	||	**2**	**26**
189	|	**1**	**27**
190		**0**	**27**
191		**0**	**27**
192	||	**2**	**29**
193	|	**1**	**30**

Ne propunem să construim un etalon în cinci clase (cvintile). Ştim că un asemenea etalon împarte lotul de subiecți în procente de câte 20%. În prima clasă se vor afla 20% dintre subiecți, cei cu înălțimea cea mai mică, urmați fiind de alți 20% dintre subiecți, cu înălțime mai mare şi aşa mai departe. În ultima clasă se vor afla 20% dintre subiecți, cu înălțimea cea mai mare.

Pentru prima clasă, vom aplica regula de trei simplă, astfel:

Dacă 30 înseamnă 100%

Atunci x înseamnă 20%

Din câte ştim, îl putem afla pe *x* înmulțind 30 cu 20 şi împărțind la 100. Deci $x=(30\times20)/100=6$.

Căutăm acum în tabel, în coloana frecvențelor cumulate, valoarea cea mai aproape de 6, valoare corespunzătoare primilor 20% dintre subiecți, cu înălțimea cea mai mică. Observăm că valoarea 6 corespunde chiar cotei 172. În prima clasă

vom include subiecții cu înălțimea cuprinsă între minim şi 172. În general, acest lucru se scrie şi astfel < 173. Acest lucru înseamnă că în prima clasă vom include toți subiecții cu o înălțime mai mică sau egală cu 172. În eşantionul normativ am avut minimul egal cu 171, aceasta fiind cea mai mică înălțime din eşantion. În practică putem oricând găsi un subiect cu o înălțime de 169. Prin urmare, prima clasă nu o vom scrie, în etalon, de forma 171 – 172 ci < 173, pentru a asigura un loc în clasificare subiecților cu valori ale caracteristicii mai mici decât minimul distribuției eşantionului normativ. Această regulă este valabilă şi pentru ultima clasă.

Pentru a doua clasă, aplicăm aceeaşi regulă de trei simplă, de data aceasta pentru un procent de 40% din eşantion:

Dacă 30 înseamnă 100%

Atunci x înseamnă 40%

Noua valoare devine 12. În tabelul frecvențelor cumulate, cel mai apropiat scor cu frecvența cumulată 12 este 175. Clasa a doua va avea, ca intervale de clasă, 173 – 175.

Tabel 6.4 – Etalon în 5 clase

Scoruri	Clasa	Semnificația
< 173	I	Foarte scunzi
173 – 175	II	Scunzi
176 – 179	III	Medii
180 – 186	IV	Înalți
> 186	V	Foarte înalți

În mod similar procedăm pentru calcularea celorlalte repere şi obținem valorile 18 pentru clasa a treia, corespunzător scorului 179, şi 24 pentru clasa a patra, corespunzător scorului 186. Având reperele claselor, putem acum scrie etalonul sub forma unui tabel ca cel de mai sus.

La distribuirea acestui etalon, nu vom uita să includem structura eşantionului normativ, distribuția scorurilor la variabila „înălțime” cu precizarea mediei, medianei, modului şi abaterii standard şi informații despre sistemul de etalonare folosit, în cazul nostru în cvintile.

Dacă aţi fost atenţi, aţi observat deja o serie de limitări ale acestui etalon. În primul rând, **intervalele inegale**. Dacă, teoretic, înălţimea minimă a unui om poate fi, să spunem, 150 de centimetri iar înălţimea maximă 220 de centimetri, atunci remarcăm numărul diferit de valori din fiecare clasă. În prima clasă putem include subiecţii de la 150 de centimetri până la 173 de centimetri, deci 23 de valori diferite. În a doua clasă includem subiecţii cu înălţimea cuprinsă între 173 şi 175, deci doar 3 valori. În clasa a treia găsim 5 valori posibile, în clasa a patra 7 valori iar în ultima clasă 33 de valori. Iată că, un asemenea etalon diferenţiază inegal subiecţii, aceasta fiind şi principala sa limită.

În al doilea rând, dacă avem o distribuţie foarte omogenă, nu mai putem diferenţia aproape deloc utilizând un asemenea etalon, deoarece intervalele de clase vor avea valori apropiate. Iată doar două motive pentru care preferăm utilizarea claselor normalizate, deoarece la nivelul acestora intervalele sunt aparent egale.

***Exerciţiu**:*

Realizaţi, folosind acelaşi sistem de etalonare, un etalon în decile (10 clase) şi precizaţi problemele întâmpinate. Explicaţi situaţia apărută.

VI.6.2 Etaloane normalizate

Această metodă de realizare a etaloanelor se bazează pe distribuţia normală şi presupune, evident, respectarea criteriului normalităţii. Dacă vom încerca să realizăm un asemenea etalon utilizând o distribuţie care nu este normală, vom avea surpriza să constatăm că anumite intervale depăşesc chiar şi amplitudinea teoretică de distribuţie a variabilei noastre. De exemplu, în cazul înălţimii subiecţilor, s-ar putea să întâlnim, în prima clasă, subiecţi cu înălţimea cuprinsă între – 30 şi – 5 centimetri ceea ce reprezintă o aberaţie.

Dacă distribuția scorurilor la variabila studiată este o distribuție gaussiană (normală), atunci putem folosi un etalon în clase normalizate, datorită avantajelor nete pe care le prezintă un asemenea sistem de etalonare.

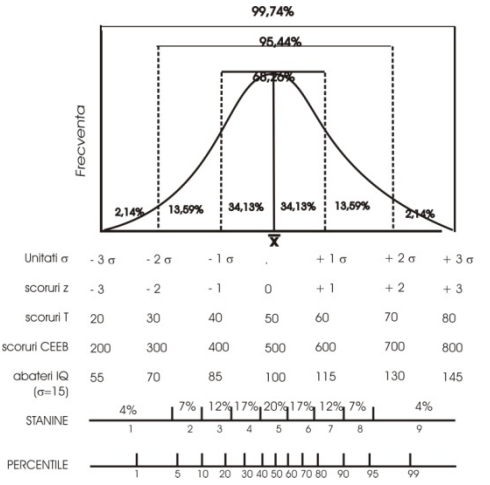


Figura 6.13 – Distribuția normală

Cel mai simplu etalon în clase normalizate este reprezentat de **unitățile sigma**, în care folosim doar media şi abaterea standard. Cu ajutorul acestor indicatori putem construi un etalon în 3, 5 şi 7 clase normalizate, luând în calcul 1, 2 sau 3 abateri standard (vezi figura alăturată). Să considerăm, de exemplu, următoarele date, reprezentând vârsta unor subiecți exprimată în ani: 35, 20, 21, 24, 29, 28, 29, 29, 24, 31, 26, 20, 38, 37, 38, 29, 24, 26, 29, 26, 24, 37, 30, 35, 24, 32, 31, 34, 39, 31. Amplitudinea de variație a acestui set de date este situată între un minim de 20 şi un maxim de 39 de ani, media este de 29,33 ani iar abaterea standard este de 5,52 ani. Aceste scoruri se distribuie normal şi putem construi un etalon în clase normalizate. Pentru a construi un etalon în cinci clase normalizate folosind unitățile sigma, va trebui să luăm intervalele formate de una şi două abateri standard în jurul mediei. Prima clasă va începe de la minim şi se va finaliza la două abateri standard în stânga mediei. Două abateri standard înseamnă 2 x 5,52 = 11,04. Dacă scădem din medie valoarea 11,04, obținem reperul pentru limita primei clase. Deci, 29,33 – 11,04 = 18,29 ani, rotunjit 18 ani. Prima clasă, care cuprinde 2,14% din populație, este situată între 0 ani şi 18 ani şi reprezintă subiecții cu vârstă foarte mică. A doua clasă se situează între 2 abateri standard şi 1 abatere standard în stânga mediei şi cuprinde 13,59% din populație. Dacă scădem din medie o abatere standard, obținem limita superioară a celei de-a doua clase. Prin urmare,

29,33 – 5,52 = 23,81 ani, rotunjit 24 de ani. A doua clasă va include subiecţii cu vârsta cuprinsă între 19 şi 24 de ani, subiecţii tineri. Clasa a treia, clasa medie, cuprinde 2x34,13% din populaţie, adică 68,26% din populaţie şi are ca limite o abatere standard în stânga mediei şi o abatere standard în dreapta mediei. Prin urmare, a treia clasă va avea ca limite de interval 25 de ani şi 29,33 + 5,52 = 34,85, rotunjit 35 de ani. Clasa a patra, care cuprinde 13,59% din populaţie, are ca limită inferioară 36 de ani şi ca limită superioară două abateri standard în dreapta mediei, 29,33 + 11,04, adică 40,37, rotunjit 40 de ani. Constatăm că deja am ieşit din amplitudinea distribuţiei noastre, ceea ce ne face să suspectăm o uşoară asimetrie. În realitate, numărul redus de subiecţi a determinat această situaţie. În sfârşit, clasa a cincia cuprinde subiecţii peste 40 de ani şi conţine 2,14% din populaţie.

Etalonul poate fi scris în acelaşi mod ca şi etalonul în cvintile şi are aceeaşi semnificaţie. Evident că, la distribuţia acestui etalon va trebui să furnizăm date legate de structura eşantionului normativ, distribuţia variabilei şi sistemul de etalonare folosit, în cazul nostru în cinci clase în unităţi sigma.

Tabel 6.5 – Etalon în 5 clase

Scoruri	Clasa	Semnificaţia
< 19	I	Foarte tineri
19 – 24	II	Tineri
25 – 35	III	Medie
36 – 40	IV	Maturi
> 40	V	Foarte maturi

***Exerciţiu**:*

Realizaţi un etalon, în unităţi sigma, în trei clase normalizate în unităţi sigma. Pentru trei clase normalizate să foloseşte o abatere standard în dreapta şi în stânga mediei.

Etaloanele în **clase normalizate** nu sunt altceva decât derivaţii ale etaloanelor în unităţi sigma şi se bazează pe transformarea notelor z în alte categorii de note. Algoritmul de calcul este oarecum similar celui utilizat în realizarea etaloanelor în unităţi sigma şi se bazează pe fracţiuni de unităţi sigma.

Astfel, un etalon în cinci clase normalizate are ca fracțiune ale lui z valoarea $^{z}/_{2}$ iar clasele devin:

I	*Min*	...	$-^{3z}/_{2}$	*6,7% din populație*
II	$-^{3z}/_{2}$	...	$-^{z}/_{2}$	*24,2% din populație*
III	$-^{z}/_{2}$	...	$^{z}/_{2}$	*38,2% din populație*
IV	$^{z}/_{2}$	...	$^{3z}/_{2}$	*24,2% din populație*
V	$^{3z}/_{2}$	...	*Max.*	*6,7% din populație*

Un etalon în şapte clase normalizate are ca fracțiuni ale lui z valoarea $^{z}/_{3}$ iar clasele devin:

I	*Min*	...	$-^{5z}/_{3}$	*4,8% din populație*
II	$-^{5z}/_{3}$	...	$-^{3z}/_{3}$	*11,1% din populație*
III	$-^{3z}/_{3}$	...	$-^{z}/_{3}$	*21,2% din populație*
IV	$-^{z}/_{3}$	...	$^{z}/_{3}$	*25,8% din populație*
V	$^{z}/_{3}$	...	$^{3z}/_{3}$	*21,2% din populație*
VI	$^{3z}/_{3}$	...	$^{5z}/_{3}$	*11,1% din populație*
VII	$^{5z}/_{4}$	...	*Max.*	*4,8% din populație*

Un etalon în nouă clase normalizate (stanine) are ca fracțiuni ale lui z valoarea $^{z}/_{4}$ iar clasele devin:

I	*Min*	...	$-^{7z}/_{4}$	*4,0% din populație*
II	$-^{7z}/_{4}$	...	$-^{5z}/_{4}$	*6,6% din populație*
III	$-^{5z}/_{4}$	...	$-^{3z}/_{4}$	*12,1% din populație*
IV	$-^{3z}/_{4}$	...	$-^{z}/_{4}$	*17,5% din populație*
V	$-^{z}/_{4}$	...	$^{z}/_{4}$	*19,6% din populație*
VI	$^{z}/_{4}$	...	$^{3z}/_{4}$	*17,5% din populație*
VII	$^{3z}/_{4}$	...	$^{5z}/_{4}$	*12,1% din populație*
VIII	$^{5z}/_{4}$	...	$^{7z}/_{4}$	*6,6% din populație*
IX	$^{7z}/_{4}$	...	*Max.*	*4,0% din populație*

În mod similar putem construi etaloane şi în alte clase normalizate luând ca referință notele z. Astfel, printre cele mai uzitate sisteme de clase normalizate putem menționa:

- **Sten (standard ten)**, distribuție cu media 2 şi abaterea standard 5,5 rezultând clase de la 1 la 10;

- **Stanford-Binet**, distribuţie cu media 100 şi abaterea standard 16;
- **Notele T**, distribuţie cu media 10 şi abaterea standard 50 rezultând clase de la 1 la 100;
- **Wechsler (abateri IQ)**, distribuţie cu media 100 şi abaterea standard 15;
- **Army General Clasification**, distribuţie cu media 100 şi abaterea standard 50;
- **Scala de aptitudini generale pentru colegiu**, distribuţie cu media 500 şi abaterea standard 100;

Tabel 6.6 – Etalon în clase normalizate

Valoare	Diagrama în linii	f	f_c
171	||	2	2
172	||||	4	6
173	|	1	7
174	||	2	9
175	|	1	10
176		0	10
177	||||	4	14
178		0	14
179	|||	3	17
180	||	2	19
181		0	19
182	|	1	20
183	|	1	21
184	||	2	23
185	|	1	24
186		0	24
187		0	24
188	||	2	26
189	|	1	27
190		0	27
191		0	27
192	||	2	29
193	|	1	30

Haideţi acum să încercăm construirea unui etalon în cinci clase normalizate folosind datele corespunzătoare înălţimii subiecţilor *182, 175, 174, 189, 177, 177, 180, 173, 188, 171, 184, 192, 180, 172, 177, 193, 184, 192, 172, 177, 174, 171, 172, 179, 185, 179, 173, 172, 179, 188.* Ne putem aminti că această distribuţie are media 183,2 centimetri cu o abatere standard de 5,23 centimetri.

Există mai multe metode de construire a unui asemenea etalon. Cea mai simplă metodă constă în utilizarea valorilor procentuale corespunzătoare fiecărei clase, într-un mod identic cu cel descris la construirea etaloanelor în cuantile.

Vom relua alăturat tabelul distribuţiei înălţimii celor 30 de subiecţi cu precizarea frecvenţelor absolute şi cumulate. Prima clasă, cea cu scorurile

cele mai mici conține primii 6,7% din eşantion. Prin regula de trei simplă, dacă 30 de subiecți înseamnă întregul eşantion, câți subiecți înseamnă 6,7%. Vom avea astfel (6,7×30)/100=210/100=2,1. La fel cum am procedat şi la etalonul în cvintile, căutăm în tabelul frecvențelor cumulate valoarea cea mai apropiată de 2,1. Această valoare este chiar prima valoare, înălțimea de 171 de centimetri. Prima clasă va cuprinde aşadar subiecții cu o înălțime mai mică sau egală cu 171 centimetri.

Următoarea clasă conține încă 24,2% din populație. Reperul de clasă va fi situat aşadar la 6,7+24,2=30,9%. În continuare aplicăm din nou regula de trei simplă. Dacă 30 de persoane înseamnă 100%, câte persoane înseamnă 30,9%. Vom avea (30,9×30)/100=9,27. Valoarea cea mai apropiată de frecvența cumulată 9,27 este 174 de centimetri. Cea de-a doua clasă va cuprinde subiecții cu înălțimea cuprinsă între 172 şi 174 de centimetri.

Reperul pentru a treia clasă va fi situat la 30,9+38,2=69,1. Analog, avem (69,1×30)/100=20,73 iar reperul de clasă va fi 183 de centimetri. A treia clasă cuprinde subiecții cu înălțimea între 175 şi 183 de centimetri.

Pentru a patra clasă vom avea 69,1+24,2=93,3 iar reperul de clasă va fi valoarea situată în dreptul frecvenței cumulate de (93,3×30)/100=27,99. Această valoare este 191 de centimetri. Clasa a patra va cuprinde subiecții cu înălțimea între 184 şi 191 de centimetri.

Evident, pentru ultima clasă nu e nevoie să mai calculăm nimic. Această clasă va conține subiecți cu înălțimea mai mare de 191 de centimetri.

Tabel 6.7 – Etalon în 5 clase

Scoruri	**Clasa**	**Semnificația**
≤ 171	I	Foarte scunzi
172 - 174	II	Scunzi
175 – 183	III	Medii
184 – 191	IV	Înalți
> 191	V	Foarte înalți

Etalonul se scrie la fel ca un etalon în cuantile şi are aceeaşi semnificație. Nu uitați să țineți cont de regulile elaborării unui etalon în momentul în care intenționați să-l folosiți sau să-l distribuiți.

VI.7 Analiza distribuţiei şi realizarea etaloanelor în SPSS

După abordarea teoretică a distribuţiei normale, a venit momentul aplicării practice a cunoştinţelor dobândite prin analiza unei distribuţii de date efectuată în SPSS. Vom utiliza aceeaşi bază de date pe care am creat-o la începutul acestui volum, baza de date *IQ.sav*.

Ne propunem să analizăm modul în care se distribuie scorurile obţinute de cei 30 de subiecţi la variabila *iq* (Coeficientul de inteligenţă).

Etapele le ştiţi deja. Vom selecta din meniul **Analyze…** opţiunea **Descriptive Statistics…** şi apoi **Frequencies…**. Apare cunoscuta casetă de dialog, de unde vom include spre analiză variabila *„iq”*.

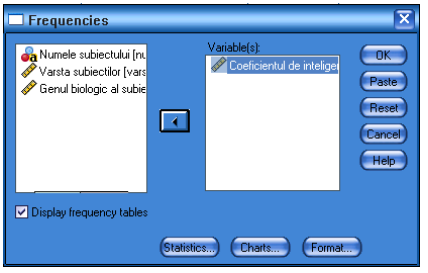


Figura 6.14 – Includerea variabilei spre analiză.

Desigur, vom accesa apoi butonul **Statistics…** pentru a comunica programului SPSS ce fel de statistici să calculeze.

Alături de indicatorii tendinţei centrale (media, mediana şi modul) şi de cei ai dispersiei (abaterea standard, amplitudinea distribuţiei, minimul şi maximul) vom include şi cei doi indicatori ai distribuţiei (Skewness şi Kurtosis), pentru a putea analiza forma acestei distribuţii şi a decide dacă distribuţia poate fi considerată sau nu o distribuţie normală.

După selectarea adecvată a acestor elemente, putem apăsa butonul **Continue** şi revenim la fereastra iniţială a selectării variabilelor. În acest moment, SPSS a memorat operaţiunile pe care trebuie să le execute.

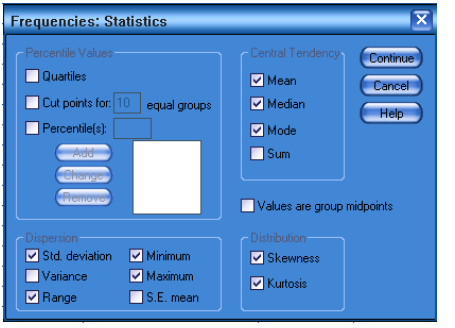


Figura 6.15 – Configurarea inventarului statistic de bază

Este uneori util să afişăm şi un grafic al distribuției datelor noastre. Pentru aceasta putem accesa butonul **Charts…** şi se va deschide fereastra de definire a graficelor pentru setul de date.

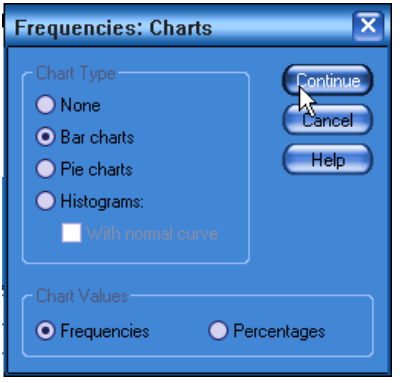


Figura 6.16 – Alegerea graficului cu bare pentru reprezentarea datelor

Această fereastră o cunoaştem destul de bine din capitolele anterioare. Motivul pentru care am abordat-o din nou este acela că, pentru a reprezenta grafic anumite date, trebuie să facem nişte precizări suplimentare.

De nenumărate ori mi-a fost dat să văd studenți care reprezintă o variabilă continuă printr-o diagramă de structură (grafic tip plăcintă), ba mai mult, profesori universitari de psihologie experimentală, coordonatori de licență, care solicită studenților reprezentarea variabilelor continui printr-un asemenea procedeu. O reprezentare prin diagramă de structură se face în cazul unor date nominale sau cel mult ordinale atunci când dorim să vedem compoziția datelor şi când numărul de valori pe care le ia variabila nu este prea mare. Este normal să reprezentăm genul biologic al subiecților printr-o plăcintă, deoarece vorbim în acest caz de X% bărbați şi Y% femei. Variabila are doar două niveluri iar această reprezentare oferă indicii clare asupra compoziției unui eşantion. Este din nou normal să reprezentăm prin plăcintă gradele didactice dintr-o universitate deoarece variabila are cinci grade de intensitate. Este însă absurd să reprezentăm grafic prin plăcintă vârsta unor subiecți, variabilă continuă, cu o infinitate de valori. Un asemenea grafic nu ne spune nimic, ci, din contra, induce distorsiuni care ne creează o imagine falsă asupra datelor.

Foarte mulți psihologi experimentaliști, care au rutină în analiza datelor, reprezintă o variabilă continuă prin histogramă, deoarece folosind acest

procedeu SPSS oferă o metodă de estimare grafică a curbei de distribuție. Unii chiar se rezumă la a interpreta distribuția datelor după aspectul grafic al acestei curbe – să fim serioşi, statistica este totuşi o disciplină matematică iar demonstrațiile se fac cu cifre, nu cu imagini şi poezii. Nu este atât de grav în comparație cu reprezentarea prin diagrame de structură, însă probabil vă amintiți de faptul că histograma reprezintă datele grupate în intervale. Este adevărat că SPSS efectuează automat o grupare a acestor date în intervale pe care le alege singur. Acesta nu este însă un motiv în condițiile în care dorim să avem o viziune analitică a scorurilor.

Prin urmare, pentru a reprezenta grafic o variabilă continuă, atunci când discutăm despre distribuție, vom prefera graficul cu bare, care ne oferă o viziune fină asupra modului în care datele se distribuie sau graficul tip histogramă, care prezintă datele într-un mod mai grosier, însă are avantajul estimării vizuale a curbei de distribuție. În nici un caz nu vom utiliza diagrame de structură sau alte tipuri de diagrame, chiar dacă acestea arată frumos.

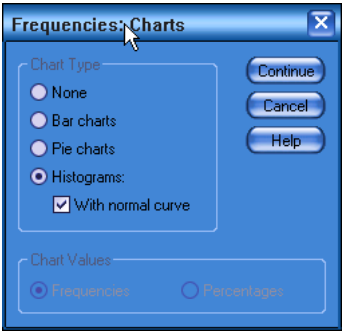


Figura 6.17 – Alegerea histogramei pentru reprezentarea datelor

În figura 6.16 am ales modul de reprezentare al unui grafic cu bare iar în figura 6.17 am optat pentru reprezentarea prin histogramă. Vom analiza ulterior ambele tipuri de grafice şi vom învăța câteva metode de „cosmetizare" a graficelor astfel încât imaginile noastre să prezinte şi un aspect elegant.

Indiferent de tipul de grafic ales (cu bare sau histogramă), vom apăsa din nou butonul **Continue** pentru a reveni la fereastra inițială. În acest moment, toate configurările fiind realizate, nu ne rămâne decât să apăsăm butonul **OK** pentru a comunica programului SPSS să efectueze analiza.

În câteva secunde va fi afişată fereastra de rezultate (Output) în care regăsim prezentate informaţiile solicitate.

Tabel 6.8 – Inventarul statistic de bază

Statistics

Coeficientul de inteligenta

N	Valid	30
	Missing	0
Mean		111,33
Median		108,50
Mode		124
Std. Deviation		8,511
Skewness		,413
Std. Error of Skewness		,427
Kurtosis		-1,384
Std. Error of Kurtosis		,833
Range		24
Minimum		100
Maximum		124

În această fereastră cel mai important tabel este tabelul **Statistics**, numit şi *inventarul statistic de bază* sau *indicatorii statistici de start*. Utilizând numai acest tabel putem decide dacă distribuţia noastră este sau nu este o distribuţie normală. Cred că vă mai amintiţi condiţiile pe care trebuie să le îndeplinească o distribuţie pentru a putea fi considerată ca fiind normală. O distribuţie normală este o distribuţie unimodală, simetrică şi mezocurtică.

Unimodalitatea este uşor de observat. SPSS raportează un mesaj de genul *„Multiple mod exist. The smallest value is shown*”, afişat la subsolul tabelului cu o notă, de obicei *a* în dreptul valorii modului. Acest mesaj ne comunică faptul că programul a identificat mai multe valori mod şi o afişează pe cea mai mică.

Tabel 6.9 – Inventarul statistic de bază

Statistics

Varsta subiectilor

N	Valid	30
	Missing	0
Mean		29,33
Median		29,00
Mode		24[a]
Std. Deviation		5,523
Skewness		,116
Std. Error of Skewness		,427
Kurtosis		-,878
Std. Error of Kurtosis		,833
Range		19
Minimum		20
Maximum		39

a. Multiple modes exist. The smallest value is shown

În inventarul statistic alăturat, unde am analizat distribuţia aceloraşi 30 de subiecţi în funcţie de variabila vârstă, avem un exemplu de distribuţie multimodală. Nu putem şti, deocamdată, dacă distribuţia este bimodală, trimodală sau prezintă

mai multe valori ale modului. Tot ceea ce ştim este că cea mai mică valoare mod, în cazul nostru, reprezintă 24 de ani.

Tabel 6.10 – Identificarea valorilor mod

Varsta subiectilor

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20	2	6,7	6,7	6,7
	21	1	3,3	3,3	10,0
	24	5	16,7	16,7	26,7
	26	3	10,0	10,0	36,7
	28	1	3,3	3,3	40,0
	29	5	16,7	16,7	56,7
	30	1	3,3	3,3	60,0
	31	3	10,0	10,0	70,0
	32	1	3,3	3,3	73,3
	34	1	3,3	3,3	76,7
	35	2	6,7	6,7	83,3
	37	2	6,7	6,7	90,0
	38	2	6,7	6,7	96,7
	39	1	3,3	3,3	100,0
	Total	30	100,0	100,0	

Pentru a afla tipul distribuţiei sub aspectul modalităţii şi a vedea toate valorile mod, vom căuta în distribuţia statistică (al doilea tabel din fereastra de rezultate) valoarea 24. În cazul nostru observăm că valoarea 24 se regăseşte în cinci cazuri (coloana **Frequency**). Nu trebuie acum decât să vedem ce valori, mai mari de 24 au aceeaşi frecvenţă. Observăm că valoarea 29 de ani are aceeaşi frecvenţă şi că nu există o altă valoare cu această frecvenţă maximă. Aşadar, distribuţia vârstei subiecţilor este o distribuţie bimodală, cu modul de 24 şi 29 de ani. Strict vorbind, o distribuţie multimodală nu este o distribuţie normală. Analiza noastră ar trebui să se oprească aici şi să folosim tehnici de eliminare a acestei anomalii (în cazul nostru renunţând la un subiect care are vârsta de 24 sau 29 de ani).

Totuşi, distribuţia scorurilor subiecţilor la variabila coeficient de inteligenţă este o distribuţie unimodală. În acest caz vom trece la următoarea etapă şi anume analiza simetriei.

Simetria sau **oblicitatea** este dată de valoarea indicatorului Skewness. La modul teoretic, acest indicator are valoarea zero pentru o distribuţie perfect simetrică. O asemenea valoare este însă aproape imposibil de regăsit în practică. Întotdeauna, acest indicator va avea valori diferite de zero, fie în sens negativ – distribuţie asimetrică negativ cu tendinţe către scoruri mari, fie în sens pozitiv – distribuţie asimetrică pozitiv cu tendinţe către scoruri mici.

Există mai multe metode de verificare a asimetriei, prin intermediul cărora putem decide dacă o distribuție poate sau nu poate fi considerată simetrică.

Una dinte metode, valabilă pentru eşantioane mari (peste 200 de subiecți) este reprezentată de reperul propus de Lewis-Beck (Sava, 2004). Valorile pentru Skewness mai mici de -0,80 sau mai mari de 0,80 indică o asimetrie clară, negativă sau pozitivă. În cazul nostru, valoarea Skewness este de 0,41 iar eşantionul este de doar 30 de persoane. Prin această metodă putem considera distribuția scorurilor ca fiind o distribuție simetrică, chiar dacă nu s-a respectat cerința unui eşantion mare.

O altă metodă (Sava, 2004) valabilă în cazul eşantioanelor medii şi mici, se bazează pe calculul cotei „z" a indicelui de oblicitate Skewness. Cu alte cuvinte, împărțim valoarea indicelui de oblicitate la valoarea erorii standard a acestuia. În acest caz, există două praguri ale lui z în funcție de valorile obținute (Sava, 2004). Pentru un eşantion mai mic de 30 de subiecți, valoarea pragului de semnificație este de 1,58 iar pentru eşantioane mai mari sau egale cu 30 de subiecți, acest prag are valoarea de 1,96. În cazul nostru, valoarea Skewness este de 0,41 iar eroarea standard Skewness este de 0,42. Făcând raportul *0,41:0,42=0,97*. Cum eşantionul nostru este de 30 de subiecți, pragul de semnificație al lui z este de 1,96. Valoarea obținută de 0,97 fiind mai mică decât pragul, putem considera distribuția ca fiind simetrică. Această metodă se foloseşte şi în cazul boltirii.

A treia metodă propusă, aplicabilă şi pentru boltire, ia în considerare stabilirea intervalului de încredere al simetriei în baza erorii standard Skewness şi poate fi utilizată indiferent de mărimea eşantionului. Eroarea standard Skewness ne oferă limitele între care trebuie să se regăsească coeficientul de oblicitate astfel încât să putem considera, la un nivel de precizie suficient de mare distribuția ca fiind simetrică. La acest nivel, există două tipuri de intervale:

- La un nivel de precizie de peste 99%, cu alte cuvinte în mai puțin de 1% din cazuri distribuția noastră poate să se comporte ca o distribuție asimetrică. Intervalul este determinat de ± o eroare standard Skewness. Un asemenea interval este foarte precis şi se foloseşte atunci când simetria reprezintă un element critic pentru distribuția datelor;
- La un nivel de precizie de peste 95%, cu alte cuvinte în mai puțin de 5% din cazuri distribuția noastră poate să se comporte ca o distribuție asimetrică. Intervalul este determinat de ± două erori standard Skewness. Un asemenea interval este mai puțin precis în comparație cu primul şi se foloseşte când simetria nu reprezintă un element critic pentru distribuția datelor.

Dacă valoarea coeficientului de oblicitate Skewness intră într-unul dintre aceste intervale, distribuția poate fi considerată simetrică. Dacă valoarea coeficientului de oblicitate iese din acest interval la limita negativă, avem de-a face cu o distribuție asimetrică negativ iar dacă iese din interval la limita pozitivă, distribuția este asimetrică pozitiv.

În situația noastră, valoarea coeficientului de oblicitate Skewness este de 0,41 iar valoarea erorii standard Skewness este de 0,42. Intervalul de încredere la un nivel de precizie de 99% este cuprins între -0,42 … +0,42 iar la un nivel de precizie de 95% este cuprins între -0,84 … +0,84. Putem constata cu uşurință că valoarea coeficientului de oblicitate Skewness (0,41) intră atât în primul interval cât şi în al doilea interval. Putem aşadar afirma, cu şanse mai mici de 1% că distribuția noastră să se comporte asimetric, că avem de a face cu o distribuție simetrică.

Indiferent de metoda folosită, am decis că distribuția scorurilor celor 30 de subiecți la coeficientul de inteligență este o distribuție simetrică. A treia metodă, datorită viziunii extrem de precise pe care o oferă, v-o recomand să o folosiți ori de câte ori analizați simetria sau boltirea unei distribuții.

Am arătat până acum că distribuția scorurilor subiecților la variabila coeficient de inteligență este o distribuție unimodală şi simetrică. Mai rămâne să demonstrăm că este o distribuție mezocurtică, pentru a decide faptul că scorurile la coeficientul de inteligență se distribuie normal.

Boltirea sau **excesul** reprezintă tendința rezultatelor de a se concentra în jurul mediei (distribuția ascuțită sau leptocurtică) sau, din contra, de a se răspândi, disipa puternic în jurul mediei (distribuția platicurtică). După cum ştim, indicatorul care ne dă mărimea acestei împrăştieri este Kurtosis. Analiza boltirii se poate face prin a doua sau a treia metodă expusă anterior, pentru care nu intrăm în detalii.

Distribuția scorurilor celor 30 de subiecți la variabila coeficient de inteligență prezintă coeficientul de boltire Kurtosis de -1,34 şi eroarea standard a acestui coeficient de 0,83.

Utilizând a doua metodă, obținem nota z pentru Kurtosis de 1,34:0,83= 1,61. Pragul de semnificație pentru eşantionul nostru este, la fel ca şi în cazul simetriei, 1,96. Valoarea notei „z" pentru Kurtosis fiind mai mică decât pragul de semnificație, putem aprecia că suntem în fața unei distribuții mezocurtice.

Haideți acum să analizăm boltirea prin prisma celei de-a treia metode pentru a vedea diferența de precizie. Intervalul de încredere la un nivel de precizie de 99% este cuprins între -0,83 … +0,83 iar pentru un nivel de precizie de 95% între -1,66 … +1,66. Observăm că valoarea Kurtosis de -1,34 intră într-al doilea interval (cu precizie mai mică) însă nu şi în primul interval (de mare precizie). Deci, în situația în care eterogenitatea sau omogenitatea scorurilor în jurul mediei este un element important, vom considera distribuția noastră ca fiind o distribuție platicurtică, deoarece nu intră în intervalul cu precizie mare. În situația în care excesul nu este un element atât de important, putem considera distribuția ca fiind mezocurtică.

Oricum, în ştiinţele sociale nivelul de precizie acceptat este de peste 95%, astfel încât distribuţia noastră poate fi considerată o distribuţie mezocurtică.

Valorile coeficientului de boltire Kurtosis negative şi care ies din intervalul de încredere sunt reprezentative pentru o distribuţie platicurtică. Dacă avem de a face cu valori pozitive care ies din intervalul de încredere, distribuţia este leptocurtică.

Demonstrând faptul că distribuţia scorurilor celor 30 de subiecţi la variabila coeficient de inteligenţă este o distribuţie unimodală, simetrică şi mezocurtică am demonstrat de fapt că distribuţia respectă toate caracteristicile unei distribuţii gaussiene şi poate fi considerată o distribuţie normală.

Analiza distribuţiei trebuie să preceadă orice analiză ulterioară de date, deoarece în funcţie de normalitatea distribuţiei pot fi alese statisticile parametrice sau nonparametrice utilizabile pe parcursul demersului de cercetare, după asumpţiile teoretice ale fiecărei metode în parte.

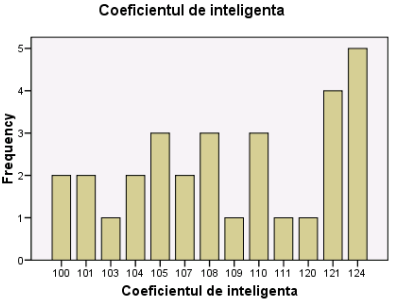
Pentru a concluziona, vă voi oferi un model complet de analiză a distribuţiei scorurilor în funcţie de variabila coeficient de inteligenţă. Întotdeauna furnizaţi, în analiză, cât mai multe informaţii legate de variabila dumneavoastră. Vă rog să aveţi în vedere ataşarea exactă a inventarului statistic de bază din SPSS şi a eventualelor grafice care vă pot susţine demonstraţia.

Într-o analiză de date, va trebui să precizaţi: denumirea variabilei analizate, numărul de cazuri, indicatorii tendinţei centrale şi ai împrăştierii, indicatorii distribuţiei şi explicarea acestora în cazul unor distribuţii asimetrice, platicurtice sau leptocurtice, tabelele valorice ale inventarului statistic de bază, graficele sub forma graficului cu bare sau histogramă.

Statistics

Coeficientul de inteligenta

N	Valid	30
	Missing	0
Mean		111,33
Median		108,50
Mode		124
Std. Deviation		8,511
Skewness		,413
Std. Error of Skewness		,427
Kurtosis		-1,384
Std. Error of Kurtosis		,833
Range		24
Minimum		100
Maximum		124



Distribuţia rezultatelor celor 30 de subiecţi la variabila „Coeficient de inteligenţă" este o distribuţie unimodală (modulul=124 IQ), simetrică (Skewness=0,41; Eroarea standard Skewness=0,42) şi mezocurtică (Kurtosis=-1,38; Eroarea standard Kurtosis=0,83), cu media 111,33 IQ, mediana 108,5 IQ şi abaterea standard 8,51 IQ. Amplitudinea distribuţiei este de 24 scoruri IQ cuprinsă între un minim de 100 IQ şi un maxim de 124 IQ.

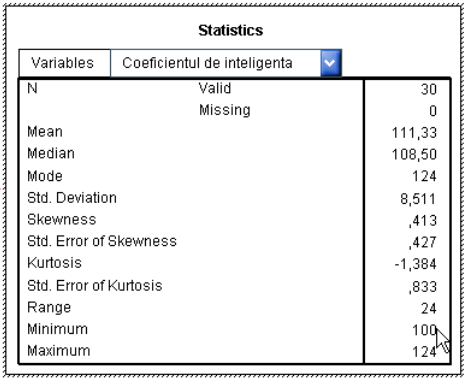
Având în vedere aceste elemente, putem considera că scorurile obţinute de cei 30 de subiecţi se distribuie normal în cazul variabilei „Coeficient de inteligenţă", această variabilă suportând proceduri parametrice de analiză statistică.

Iată un exemplu complet şi profesional prin care puteţi realiza o analiză a distribuţiei datelor, validă în orice comunicare ştiinţifică naţională sau internaţională. Mulţi vor spune că nu e chiar profesional, că trebuia să specific tabelul numărul …. figura numărul…. şi ce înseamnă acel tabel şi acea figură. Asemenea lucruri ţin de normele de redactare ale documentelor ştiinţifice elaborate APA şi vi le las ca studiu, deoarece nu intenţionez să le abordez în prezenta lucrare. Oricum, mă adresez unor oameni inteligenţi iar aceste amănunte le ştiţi sau le puteţi afla şi singuri. Nici cartea de faţă nu le respectă, însă prefer să am un stil propriu şi aş fi încântat dacă aţi învăţa din

aceste rânduri **cum** să faceţi efectiv analize de date mai mult decât **cum** să vă conformaţi unor norme.

Ca o paranteză, desigur am să fiu acuzat că am copiat şi lipit în document tabele şi grafice din SPSS fără să le traduc în româneşte. Departe de mine această intenţie. În primul rând, pentru că SPSS le oferă în engleză şi doresc să introduc conţinutul original. În al doilea rând, conform aceloraşi norme APA, utilizarea elementelor furnizate de programe informatice în lucrări ştiinţifice se face prin includerea conţinutului original şi nemodificat. În al treilea rând, pentru că vreau să vă învăţ, imediat, cum să editaţi conţinutul elementelor din fereastra de rezultate.

Orice element din fereastra de rezultate SPSS este un element editabil. Puteţi să faceţi aproape orice cu tabelele şi graficele pe care SPSS le furnizează. Nu aveţi decât să daţi un dublu clic pe elementul pe care doriţi să-l editaţi şi veţi intra într-un alt mod pe care SPSS îl suportă, modul de editare.



Statistics

Variables: Coeficientul de inteligenta

N	Valid	30
	Missing	0
Mean		111,33
Median		108,50
Mode		124
Std. Deviation		8,511
Skewness		,413
Std. Error of Skewness		,427
Kurtosis		-1,384
Std. Error of Kurtosis		,833
Range		24
Minimum		100
Maximum		124

Figura 6.18 – Tabelul Statistics în mod editare

Pentru început să dăm dublu clic pe tabelul inventarului statistic de bază, tabelul Statistics. Observăm că marginea care încadrează acest tabel a devenit o margine haşurată şi de asemenea este posibil să se afişeze o bară de instrumente de editare, similară celei din programul Microsoft Word.

Primul buton de pe această bară de formatare permite anularea acţiunilor efectuate (Undo). Cel de-al doilea buton este butonul „pivot” ce permite schimbarea rândurilor şi a coloanelor în tabel.

Figura 6.19 – Bara de formatare SPSS

Urmează apoi o casetă derulantă de aplicare a unui stil tabelului formatat, caseta derulantă de alegere a fontului şi caseta derulantă de alegere a dimensiunii fontului pentru textele selectate. Următoarele trei butoane permit scrierea îngroşată, înclinată şi sublinierea iar butonul al patrulea selectează culoarea cu care să fie afişat fontul.

Cele patru butoane care se succed permit alinierea textului selectat, la stânga, dreapta, centrat sau automat.

Ultimele două butoane lansează sistemul de asistenţă (help) sau modul de creare a graficelor pentru variabila selectată – iată o altă modalitate de a realiza grafice dacă ați uitat să o faceți în etapa de analiză. Vă recomand să exersați aceste elemente, fără teamă că ați putea strica ceva (în definitiv aveți la dispoziție butonul de anulare - Undo). Nu vom intra în amănunte legate de formatarea tabelelor deoarece ar trebui să includem, inutil, mai multe pagini.

Statistics

Variables	Coeficientul de inteligenta	
N	Valid	30
	Missing	0
Mean		111,33
Median		108,50
Mode		124
Std. Deviation		8,511
Skewness		,413
Std. Error of Skewness		,427
Kurtosis		-1,384
Std. Error of Kurtosis		,833
Range		24
Minimum		100
Maximum		124

Figura 6.20 – Editarea valorilor în tabel

Un tabel în modul de editare permite modificarea elementelor componente. Pentru început ne propunem să traducem în românește cuvintele Missing (Lipsă), Mean (Media), Media (Mediana) şi Mod (Modul).

Aflându-vă în modul de editare, efectuăm din nou dublu clic pe cuvântul Missing. Va rezulta o figură similară celei alătu-

rate. Observați că textul Missing este deja selectat acum; dacă apăsați orice tastă, conținutul selectat va fi înlocuit cu tasta apăsată. Nu vă rămâne decât să scrieți în loc de Missing, Lipsă.

Cuvântul „Missing" este înlocuit cu textul introdus, „Lipsă". Procedați la fel cu toate celelalte cuvinte din tabel, pentru a putea traduce tabelul în limba română. Puteți folosi bara de formatare pentru a îngroşa, schimba culoarea sau alinia textele selectate. Când ați terminat, dați clic oriunde pe suprafața albă a ferestrei de rezultate pentru a ieşi din modul de editare.

Inventar statistic de baza

Coeficientul de inteligenta

N	Valid	30
	Lipsa	0
Media		111,33
Mediana		108,50
Modulul		124
Abaterea standard		8,511
Skewness		,413
Eroarea Skewness		,427
Kurtosis		-1,384
Eroarea Kurtosis		,833
Amplitudinea		24
Minimum		100
Maximum		124

Figura 6.21 – Inventarul statistic după editare

În final, tabelul dumneavoastră trebuie să arate la fel ca cel din figura de mai sus.

Iată modalitatea prin care puteți traduce dacă doriți orice element al formularului de rezultate într-o altă limbă.

Din nefericire, la fel cum pot fi traduse cuvintele, pot fi modificate şi cifrele, ceea ce vă poate determina uneori să schimbați datele oferite de SPSS cu alte date mai convenabile vouă. Nu uitați că în general trebuie să puneți la dispoziție şi baza de date în momentul în care publicați un studiu şi de cele mai multe ori analizele sunt refăcute de comisia de publicare a articolului.

Tot folosind modul de editare, puteți ajusta lățimea coloanelor pentru ca tabelul dumneavoastră să prezinte un aspect elegant.

Deplasați cursorul mausului deasupra unei linii, ca în figura alăturată, până când se transformă într-o săgeată orizontală. Apoi țineți apăsat butonul din stânga al mausului şi „trageți" mişcând mausul la dreapta sau la stânga

Inventar statistic de baza

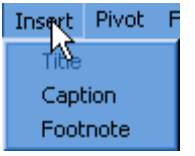
Variables	Coeficientul de inteligenta	
N	Valid	30
	Lipsa	0
Media		111,33
Mediana		108,50
Modulul		124
Abaterea standard		8,511
Skewness		,413
Eroarea Skewness		,427
Kurtosis		-1,384
Eroarea Kurtosis		,833
Amplitudinea		24
Minimum		100
Maximum		124

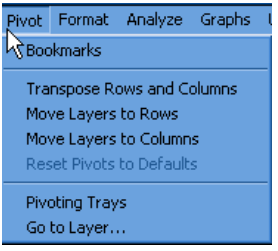
Figura 6.22 – Inventar statistic tradus

pentru a ajusta dimensiunea coloanei din tabel. La fel puteți proceda și pentru ajustarea dimensiunilor pe verticală.

Referitor la formatarea tabelelor, SPSS pune la dispoziție mai multe meniuri pe care nu le vom analiza în detaliu aici din considerente de spațiu. Puteți consulta, pentru informații suplimentare, manualul SPSS. Ne vom limita doar să le prezentăm.

Meniul **Insert** permite adăugarea, în cadrul tabelului selectat, a unui nou titlu, a unei noi etichete în subsolul tabelului sau a unei noi note se subsol într-un mod similar operațiunii pe care SPSS o face automat în momentul în care depistează o distribuție multimodală.

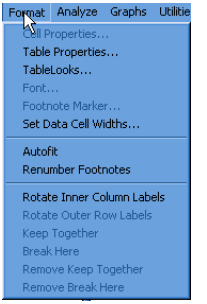




Meniul **Pivot** permite marcarea tabelului curent în scopul unei regăsiri rapide în fereastra de rezultate, comutarea tabelului prin inversarea liniilor sau a coloanelor sau lansarea ferestrei speciale prin care poate fi controlată pivotarea tabelului. Acest meniu este un meniu important în formatarea aspectului grafic al tabelului, cu aplicații în condițiile analizei unui volum mare de date. Vom reveni asupra acestui meniu în momentul în care vom discuta despre studiile corelaționale.

Meniul **Format** cuprinde un set de utilitare prin intermediul cărora putem modifica aspectul celulelor din tabel, a întregului tabel, a fonturilor

sau a notelor de subsol. SPSS conţine, asemănător programului Microsoft Word, mai multe şabloane grafice ce pot fi aplicate tabelelor. Alegând un asemenea şablon grafic, puteţi modifica complet aspectul tabelelor din foaia de rezultate.



Toate opţiunile din acest meniu ţin doar de aspectul grafic al tabelelor şi nu afectează în nici un fel conţinutul datelor dumneavoastră. Se pot stabili, alături de cele expuse mai sus, numărul de zecimale pe care SPSS să le afişeze, ordonarea datelor, includerea sau excluderea capului de tabel etc. Vă invit să parcurgeţi acest meniu şi să exersaţi facilităţile pe care le oferă. Oricând vă puteţi întoarce la starea iniţială, anterioară unei acţiuni, folosind butonul Undo.

VI.7.1 Calculul notelor „z" în SPSS

Programul SPSS permite calculul notelor „z" pentru fiecare dintre cazurile individuale prezente în distribuţia de date, într-un mod foarte simplu şi elegant.

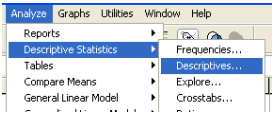


Figura 6.23 – Meniul statistici descriptive

Din acelaşi meniu, **Analyze**, veţi alege opţiunea **Descriptive Statistics** şi apoi opţiunea **Descriptives…**

Cunoaşteţi deja fereastra care se deschide, deoarece am analizat-o pe parcursul acestei lucrări, drept pentru care nu vom insista decât asupra unui singur element. Este vorba despre caseta de bifare **„Save standardized values as variables".** Rolul acestei casete, pe care o vom bifa, este acela de a crea o nouă variabilă ce va stoca scorurile „z" ale fiecărui caz din baza de date. Evident, va fi lansată fereastra de afişare a rezultatelor ce conţine un tabel pe

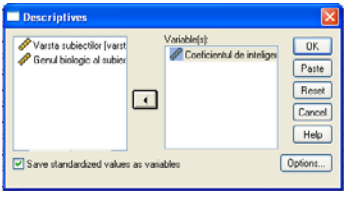


Figura 6.24 – Fereastra de configurare a analizei

care l-am discutat deja. Totuşi, unde sunt notele „z” atât de căutate de noi?

	sexul	iq	Ziq
35	Masculin	100	-1,33161
20	Feminin	120	1,01829
21	Feminin	124	1,48827
24	Masculin	121	1,13579
29	Feminin	124	1,48827
28	Masculin	101	-1,21412
29	Feminin	110	-,15666
29	Masculin	124	1,48827
24	Feminin	121	1,13579
31	Feminin	121	1,13579
26	Masculin	100	-1,33161
20	Masculin	101	-1,21412
38	Masculin	105	-,74414
37	Masculin	108	-,39165
38	Feminin	104	-,86163
29	Feminin	110	-,15666

Figura 6.25 – Noua variabilă ce conţine scorurile z

Să ne amintim ce am bifat. Am bifat o opţiune după ce am introdus variabila „Coeficient de inteligenţă” în lista variabilelor ce trebuie analizate, prin care notele standardizate vor fi salvate ca variabile. Trebuie deci să ne întoarcem la baza de date, unde vom regăsi o nouă variabilă creată automat. Este vorba despre variabila „Ziq”, ultima din baza de date, care nu conţine altceva decât notele „z” ale fiecărui subiect.

Observăm că primul subiect se află la 1,33 abateri standard în stânga mediei, adică în zona rezultatelor medii, al doilea la 1,01 abateri standard în dreapta mediei, tot în zona scorurilor medii şi aşa mai departe.

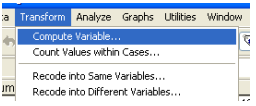


Figura 6.26 – Meniul de transformare SPSS

Cum facem totuşi pentru a obţine note standardizate rezultate din notele „z”, scoruri sten spre exemplu. Vă mai amintiţi probabil că scorurile sten se obţin adunând valoarea 5,5 la valoarea notei „z”. Dacă avem notele „z”, restul este simplu. Cum facem însă acest lucru în SPSS?

Vom folosi serviciile unui alt meniu şi anume meniul **Transform,** de unde vom selecta opţiunea **Compute** sau **Compute variable…**. în funcţie de versiunea SPSS pe care o posedaţi.

Rezultatul acestei acţiuni se concretizează într-o fereastră similară celei alăturate, care conţine următoarele elemente:

Figura 6.27 – Fereastra de calcul a variabilelor

Target variable reprezintă secţiunea în care vom introduce numele variabilei destinaţie, variabilă

în care se vor salva noile rezultate. În cazul nostru, suntem interesaţi să calculăm scorurile sten pentru variabila „iq" şi vom introduce un nume, ca spre exemplu „StenIQ".

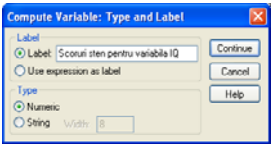


Figura 6.28– Proprietăţi ale variabilei

Sub această casetă de text se află un buton intitulat **Type & Label…** pe care dacă apăsăm vom putea configura proprietăţile noii variabile create.

Astfel, avem posibilitatea să adăugăm o etichetă variabilei nou create folosind cele două opţiuni din secţiunea **Label**. Putem decide dacă introducem un text ca şi etichetă (ca în cazul figurii alăturate) ori vom folosi drept etichetă formula de calcul, situaţie în care va trebui să alegem opţiunea **Use expression as label**. Secţiunea **Type** permite alegerea tipului de variabilă. Putem decide între un tip numeric implicit sau între un şir de caractere, caz în care va trebui să includem şi dimensiunea acestui şir, numărul de caractere pe care îl permite variabila.

Sub butonul Type & Label se află binecunoscuta listă a variabilelor din baza de date. În partea dreaptă apare caseta de text **Numeric Expression** cu ajutorul căreia putem scrie formula de calcul prin care va fi obţinută noua variabilă. În acest sens putem utiliza „calculatorul" de sub casetă sau putem tasta pur şi simplu valorile sau formula în caseta de text. Pentru a vă uşura accesul la formule, SPSS vă pune la dispoziţie două liste: lista **Function group** care conţine funcţiile sistemului grupate pe categorii şi lista **Functions and Special Variables** care se referă la funcţiile din categoria selectată anterior. Dacă, de exemplu, în prima listă efectuăm clic pe elementul **Arithmetic**, în cea de-a doua listă vor fi afişate funcţiile referitoare la operaţiile aritmetice uzuale.

Butonul **If…** permite efectuarea condiţională a calculelor, adică calcularea doar a datelor care îndeplinesc o condiţie.

Întâlnim aici posibilitatea de a include toate cazurile (**Include all cases**) sau a include doar cazurile care satisfac o condiție (**Include if cases satisfies condition**). Dacă, de exemplu, dorim să calculăm scorurile sten doar pentru bărbați, vom alege această opțiune iar apoi vom scrie în caseta de formule expresia „sexul=1”. În acest fel SPSS va calcula scorurile sten numai pentru bărbați, ignorând femeile. Întâlnim şi aici lista de funcții, utilă pentru crearea unor expresii mai complexe. Deocamdată nu ne interesează nici un calcul condițional, prin urmare vom părăsi fereastra fără a efectua nici o modificare.

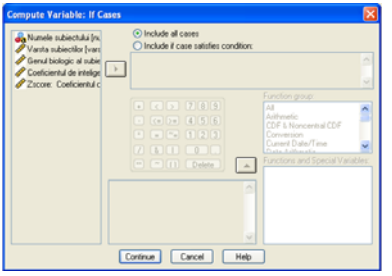


Figura 6.29 – Calculul condițional al datelor

Să vedem acum, concret, cum calculăm scorurile sten ale variabilei „iq” folosind aceste elemente. În primul rând, am introdus numele şi descrierea noii variabile în caseta Target Variable, aşa cum a fost exemplificat mai sus. Ne amintim că nota sten se calculează adunând 5,5 la valoarea notei „z”. Nota „z” însă am calculat-o anterior şi o regăsim în baza de date sub denumirea „Ziq (Zscore: Coeficientul de inteligență).

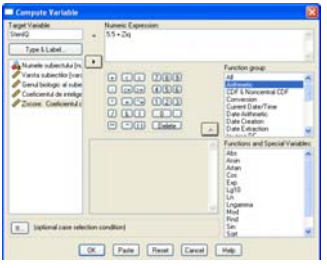


Figura 6.30 – Calculul scorurilor sten

Restul e simplu. Nu vă rămâne decât să scrieți în caseta Numeric Expression formula (5,5 + Ziq), fie tastând-o, fie folosind „calculatorul”. Variabila care conține nota „z” (Ziq) se poate tasta sau se poate folosi butonul de transfer pentru a include variabila din lista variabilelor în formulă.

Nu mai trebuie decât să apăsați butonului **OK** şi să vedem ce se întâmplă. Apare, desigur, fereastra de rezultate care ne informează asupra faptului că operațiunea a reuşit. La fel ca la calculul notelor „z”, ne interesează baza de date.

Iată că a apărut noua variabilă „StenIQ” care conţine, exact aşa cum am specificat, scorurile sten ale subiecţilor. Observăm că de această dată notele nu mai sunt cuprinse între -3 şi +3 ci pe o scală de la 1 la 10, exact ca în sistemul şcolar. Este mult mai intuitiv, nu-i aşa?

Ca exerciţiu, vă propun calculul notelor „z” şi a scorurilor sten pentru variabila „vârsta subiecţilor”.

Concluzii:

- *Media şi abaterea standard caracterizează pe deplin o distribuţie şi de aceea aceşti doi indicatori poartă numele de parametri ai repartiţiei normale;*
- *O distribuţie normală trebuie să îndeplinească simultan următoarele trei condiţii:*
  - *Să fie unimodală;*
  - *Să fie simetrică;*
  - *Să fie mezocurtică;*
- *Unimodalitatea reprezintă existenţa unei singure categorii cu frecvenţa absolută maximă şi se analizează prin inspectarea valorii modale;*
- *O distribuţie normală este o distribuţie unimodală;*
- *Simetria reprezintă echilibrarea distribuţiei pe axa orizontală şi se poate calcula prin:*
  - *Coeficientul Yule;*
  - *Coeficientul Fisher;*
  - *Relaţia dintre medie, mediană şi abaterea standard;*
- *Referitor la simetrie există un sistem de relaţii între indicatorii tendinţei centrale (medie, mediană şi mod), distribuţiile putând fi simetrice, asimetrice la stânga şi asimetrice la dreapta.*
- *O distribuţie normală este o distribuţie simetrică*
- *Boltirea reprezintă o „asimetrie verticală” a distribuţiei şi se poate calcula prin:*
  - *Coeficientul de boltire Pearson;*
- *O distribuţie normală este o distribuţie mezocurtică;*
- *Într-o distribuţie normală există un număr de şase abateri standard; trei la stânga mediei şi trei la dreapta mediei;*
- *În funcţie de abaterile standard, o distribuţie normală prezintă patru zone: zona scorurilor normale, zona scorurilor accentuate, zona scorurilor atipice şi zona scorurilor aberante;*
- *Notele „z” reprezintă distanţa dintre un scor particular şi medie exprimată în termeni de abateri standard;*

- *Notele „z” pot fi transformate, în baza proprietăților acestora, în alte sisteme de note standardizate;*
- *Prin intermediul notelor „z”, scorurile brute se aduc la un numitor comun, putând apoi face comparații atât între subiecți cât şi între rezultate la probe diferite;*
- *O distribuție care nu este normală poate fi normalizată prin procedee statistice specifice:*
    - *Media 5% trim;*
    - *Extragerea radicalului;*
    - *Ridicarea la pătrat sau la cub;*
    - *Logaritmarea pe valori normale sau inversate;*
    - *Reflectarea.*
- *Etaloanele sunt sisteme de norme ce permit compararea scorului unui subiect cu performanța obținută de un grup reprezentativ de subiecți;*
- *Etaloanele se construiesc în funcție de aspectul distribuției rezultatelor, astfel:*
    - *Pentru distribuții normale se pot folosi etaloanele în cuantile sau etaloane în clase normalizate, preferându-se cele din urmă;*
    - *Pentru distribuții care nu sunt normale şi nu pot fi normalizate, singurul sistem de etalonare care poate fi folosit este cel al etaloanelor în cuantile;*

VII. PROBABILITĂŢI ŞI SEMNIFICAŢIE STATISTICĂ

În acest capitol se va discuta despre:

- Conceptul de probabilitate;
- Tipuri de probabilităţi;
- Eroare standard şi intervale de încredere;
- Etapele metodei ştiinţifice;
- Ipoteze şi testarea ipotezelor;
- Semnificaţie statistică şi erori în testarea ipotezelor.

După parcurgerea capitolului, cursanţii vor fi capabili să:

- Înţeleagă regulile de probabilitate;
- Distingă între diferitele tipuri de probabilităţi;
- Înţeleagă şi să calculeze eroarea standard şi intervalul de încredere;
- Însuşească etapele demersului ştiinţific;
- Planifice o cercetare ştiinţifică şi să formuleze ipotezele;
- Înţeleagă logica testării ipotezelor;
- Înţeleagă conceptul de semnificaţie statistică;
- Recunoască şi să trateze tipurile de erori apărute în procesul testării ipotezelor.

Desigur, în activitatea dumneavoastră de zi cu zi, aţi întâlnit frecvent termenul de probabilitate. Care este probabilitatea să plouă în momentul în care aţi decis să vă petreceţi o săptămână pe litoral sau care este probabilitatea să aveţi un cămin fericit în momentul în care v-aţi căsătorit? Dacă aruncaţi o monedă în sus, care este probabilitatea să cadă cap sau stemă? Unii dintre dumneavoastră îşi doresc să poată afla probabilitatea prin care să obţină un şase la zar pentru a scăpa de marţ şi aşa mai departe.

Dacă aruncăm o monedă în sus, atunci când cade, va putea să cadă pe faţa cu capul sau pe faţa cu stema. Există doar două posibilităţi în acest caz care acoperă complet posibilităţile de manifestare ale acestui fenomen. Care este însă probabilitatea să cadă stema? Este una din cele două posibilităţi,

adică exprimat în termeni cifrici ½=0,5. Dacă avem un zar, care este probabilitatea să dăm un şase? De data aceasta câte posibilităţi sunt? Poate să cadă 1, 2, 3, 4, 5 sau 6, deci avem 6 posibilităţi. Dacă dăm cu zarul, există o singură şansă raportată la 6 şanse de a nimeri, dintr-o singură încercare, zarul cu şase. Exprimat din nou cifric, avem o probabilitate de $^1/_6$=0,166 să scăpăm de marţ. Dacă însă pentru a scăpa de marţ ne trebuie un 6 sau un 4, care este probabilitatea de a ieşi cu faţă curată? De data aceasta există două alternative raportate la un total de şase, prin urmare $^2/_6$=0,33.

Nu cred că toate aceste lucruri sunt elemente noi. În general, atunci când vorbim de probabilităţi, acestea se exprimă sub formă cifrică, cu valori de la 0 la 1, în care 0 reprezintă probabilitatea nulă, niciodată evenimentul nu se va întâmpla iar 1 reprezintă probabilitatea maximă, evenimentul se întâmplă întotdeauna. De foarte multe ori este mai intuitiv să reprezentăm procentual probabilităţile prin înmulţirea probabilităţii cu 100. Astfel, există 50% şanse să cadă stema la aruncarea unei monede (0,5x100) sau 16,6% şanse să scap de marţ dacă nimeresc un şase şi 33% dacă nimeresc un şase sau un patru. Mulţi preferă această exprimare deoarece suntem obişnuiţi să gândim sub formă de părţi ale unui întreg. În definitiv, fie că folosim exprimarea 0,33 fie că avem în vedere o formă procentuală – 33% – de fapt spunem acelaşi lucru.

Toate aceste exemple reprezintă cazuri de **probabilitate necondiţionată,** deoarece nu există nici un factor, nici un element care să influenţeze desfăşurarea lor. Dacă nu există vreun magnet în zar sau vreun grăunte de plumb care să influenţeze obţinerea lui şase, atunci suntem în faţa unui caz de probabilitate necondiţionată.

Probabil că la ora actuală există o sută de cărţi de statistică care tratează problemele pe care le-am expus aici, unele în mod cert mai complete şi mai academic scrise. Care ar fi aşadar probabilitatea ca dumneavoastră să citiţi acum cartea scrisă de mine? Este simplu de calculat $^1/_{100}$=0,01 sau exprimată procentual 1%. Cam mici şanse, nu-i aşa? În mod sigur nu cred c-am

să mă îmbogăţesc din vânzarea acestei cărţi… Care ar fi probabilitatea ca prietenul sau prietena dumneavoastră să citească această carte? Tot 1% dacă nu i-o recomandaţi dumneavoastră. Dacă i-o recomandaţi, probabilitatea va creşte în funcţie de numărul de recomandări efectuate şi, iată cum cu ajutorul dumneavoastră îmi voi lua o vilă pe plajă. Acest exemplu, în care probabilitatea de apariţie a unui eveniment este influenţată de prezenţa sau absenţa unui alt eveniment se numeşte **probabilitate condiţionată**. Există aşadar alte aspecte care influenţează probabilitatea de apariţie a unui eveniment. Probabilitatea de a muri de cancer la plămâni în urma fumatului este foarte mare în condiţiile în care nu cunosc faptul că fumatul îmi poate declanşa un cancer. Dacă ştiu acest lucru, creşte probabilitatea să reduc ţigările şi chiar să abandonez fumatul, ceea ce va determina şi reducerea probabilităţii de a muri. În prima situaţie ne aflăm în faţa unui exemplu de probabilitate necondiţionată. În al doilea caz, avem de a face cu o probabilitate condiţionată.

De ce totuşi am abordat aceste probleme pe care le ştiţi atât de bine? Nu am făcut-o ca să va jignesc inteligenţa, ci pentru a ne asigura de bazele înţelegerii importanţei probabilităţilor în statistica socială.

Înţelegerea acestor aspecte asigură, mai apoi, înţelegerea conceptelor de inferenţă statistică, de relaţii între variabile, praguri de semnificaţie şi semnificaţie statistică.

După cum am afirmat, scopul unei cercetări este acela ca, pornind de la rezultatele obţinute la nivelul unui eşantion de populaţie să tragem concluzii valabile pentru o întreagă populaţie. Dacă voi studia cantitatea de bere consumată de studenţii de la psihologie şi voi face afirmaţia că în general studenţii de la psihologie consumă mai multă bere în comparaţie cu ceilalţi studenţi, la ce nivel de precizie mă voi situa? Care este probabilitatea ca afirmaţia mea să fie exactă? Pentru a calcula această probabilitate avem nevoie să înţelegem foarte clar ce înseamnă probabilitate de apariţie a unui eveniment, fie ea condiţionată sau necondiţionată. Toate cercetările în care se folo-

sesc procedee statistice urmăresc, în principal, acelaşi scop. Din raţiuni de costuri şi volum de muncă, studiind un eşantion, urmărim extinderea concluziilor obţinute la nivelul întregii populaţii. Singura problemă care se pune este cea referitoare la cât de precise şi adecvate sunt aceste concluzii.

Probabilitatea condiţionată şi simplă necondiţionată reprezintă două dintre cele mai importante reguli de probabilitate, alături de evenimentele mutual exclusive şi evenimentele independente. În următoarele subcapitole vom aprofunda aceste noţiuni în vederea unei înţelegeri complete.

VII.1Patru reguli de probabilitate

În exemplul anterior cu zarul, totalitatea rezultatelor care pot să fie obţinute (1, 2, 3, 4, 5 sau 6) poartă numele de **eveniment** în teoria probabilităţilor. Probabilitatea de scăpa de marţ dacă dau un şase sau dacă dau un şase sau un patru, poartă numele de probabilitate de succes. Generalizând, probabilitatea de succes în cazul unor asemenea evenimente poate fi calculată cu ajutorul formulei $P_{(s)} = \frac{n_s}{n}$, în care $P_{(S)}$ reprezintă probabilitatea de succes, n_s reprezintă numărul rezultatelor de succes iar n reprezintă numărul total al rezultatelor.

Este exact ceea ce am stabilit anterior. Care este probabilitatea de a scăpa de marţ (probabilitatea de succes)? Să dau un şase. Deci un singur rezultat va fi de succes dintr-un total de câte rezultate? De şase rezultate posibile (pot să dau un 1, 2, 3, 4, 5 sau 6). Aplicând în formulă, găsim că probabilitatea de succes este de $^1/_6$=0,166.

În mod similar pot calcula probabilitatea de succes pentru două evenimente (şase sau patru), pentru trei evenimente şi aşa mai departe.

VII.1.1 Probabilitatea simplă pentru evenimente egale

Se referă exact la evenimente tip aruncarea unei monede sau la aruncarea unui zar şi reprezintă cea mai simplă regulă de probabilitate. Pentru a

înțelege mai bine acest lucru, să luăm cazul unei rulete cu nouă culori. Dumneavoastră va trebui să aruncați o monedă de 50 de bani în timp ce ruleta se roteşte. Dacă atunci când ruleta se opreşte moneda dumneavoastră se află pe culoarea verde, câştigați doi lei; în caz contrar, pierdeți moneda.

Cunoscând formula de mai sus, care este probabilitatea de succes? Evident, numărul total de rezultate posibile este nouă; moneda poate să cadă pe oricare dintre cele nouă culori. O singură culoare este însă câştigătoare şi anume culoarea verde, deci un singur rezultat va fi cel câştigător. Probabilitatea de succes este aşadar $^1/_9$=0,111 sau 11,1%. Până aici toate sunt clare şi limpezi. Care este însă probabilitatea de eşec? Evident toate celelalte culori rămase, adică $^8/_9$=0,888 sau 88,8%. Din nouă şanse posibile, aveți aşadar o singură şansă să câştigați şi opt şanse să pierdeți, adică $^1/_8$ şanse de câştig. Acum vă veți uita nedumeriți la mine şi mă veți întreba…. Parcă era $^1/_9$ probabilitatea de succes. Cum a devenit dintr-o dată $^1/_8$? Dacă ați fost atenți, am făcut deja incursiunea într-un alt concept statistic şi anume cel de **şansă**, oarecum similar cu cel de probabilitate de succes şi de eşec, pe care de altfel se bazează. **Şansa de succes** poate fi scrisă ca raportul dintre probabilitatea de succes şi probabilitatea de eşec ($^{P(s)}/_{P(e)}$). Pentru a nu mai folosi probabilitatea de eşec, putem exprima şansa de succes doar în termeni de probabilitate de succes, după formula $Sansa = \frac{P_{(s)}}{1-P_{(s)}}$. Acest lucru devine evident dacă înlocuim în raportul inițial semnificația probabilității de eşec. Atunci când vorbim în termeni de şansă, putem spune că există o şansă să câştig doi lei şi opt şanse să pierd 50 de bani, adică şansa mea de succes este de $^1/_8$; una de câştig şi opt de pierdere.

Haideți să vedem dacă aşa stau lucrurile. Probabilitatea de succes, în cazul ruletei, este de $^1/_9$ şi cea de eşec este de $^8/_9$. Vom face apel la cunoştințe de aritmetică de bază şi vom spune că şansa de succes este $\frac{P_{(s)}}{P_{(e)}} = \frac{^1/_9}{^8/_9} = \frac{1}{9} \times \frac{9}{8} = \frac{1}{8}$. Nu am folosit decât cunoştințe elementare de aritmetică. Demonstrația

formulei a doua este evidentă deoarece probabilitatea de eşec va fi 1 – probabilitatea de succes. Dacă înlocuiți şi efectuați calculele, rezultatul va fi acelaşi: o şansă să câştig doi lei şi opt şanse să pierd 50 de bani.

De ce folosim totuşi termenul de şansă de succes şi nu cel de probabilități de succes şi de eşec. Foarte simplu. Pentru a sintetiza într-o singură exprimare atât succesul cât şi eşecul. „O şansă să câştig doi lei şi opt şanse să pierd 50 de bani" se prezintă mai concis şi mai elegant în comparație cu o probabilitate de unu pe nouă să câştig doi lei şi o probabilitate de opt pe nouă să pierd 50 de bani. Desigur, sunt şi alte motive pe care le vom discuta la momentul potrivit.

VII.1.2 Evenimentele mutual exclusive

Toată discuția noastră s-a bazat până acum pe un singur eveniment: aruncarea unui singur zar, a unei singure monede sau jocul la o singură ruletă. Ce se întâmplă însă dacă avem două evenimente? Ştiți că la jocul de table avem două zaruri. Fiecare dintre cele două zaruri poate lua, în mod independent, valori de la unu la şase în urma unei aruncări. Dacă pentru a scăpa de marț nu avem nevoie de un şase sau un patru ci de un şapte sau un zece, cum vom calcula probabilitatea? Să nu îmi spuneți că veți picta mai multe puncte pe fața unui zar…

Iată o situație tipică de două evenimente mutual exclusive. De ce mutual exclusive? Pentru că punctele obținute pe un zar depind de punctele obținute pe celalalt zar pentru a rezulta scorul total. Fiecare dintre cele două zaruri poate lua valori de la unu la şase. Cumulat, există un număr de 36 de posibilități de a se combina cele două zaruri (6x6 posibilități). Pentru a obține un şapte, putem avea una dintre următoarele situații:

Zar 1:	*1*	*2*	*3*	*4*	*5*	*6*
Zar 2:	*6*	*5*	*4*	*3*	*2*	*1*
Total:	*7*	*7*	*7*	*7*	*7*	*7*

În acest caz funcționează o altă regulă, derivată din regula probabilităților simple a evenimentelor egale, numită regula **aditivă pentru evenimente mutual exclusive.** Nu vă speriați că nu este cine ştie ce demonstrație matematică complexă. Care este numărul total de rezultate al celor două evenimente? Primul zar are un total de şase rezultate iar al doilea zar un număr total de şase rezultate. Numărul total al rezultatelor celor două evenimente este de 6x6=36 de rezultate. Care este numărul de rezultate de succes, adică numărul de rezultate care prin însumare duc la cifra şapte? Numărați rezultatele din tabelul de mai sus. Sunt şase rezultate. Nu avem acum decât să înlocuim în formula de mai sus pentru a calcula probabilitatea de succes $P_{(s)} = \frac{n_s}{n} = \frac{6}{36} = \frac{1}{6} = 0{,}16$. Cunoscând probabilitatea de succes este uşor să calculăm şansa de succes. Aplicați doar formula specificată în capitolul anterior şi veți afla şansa de succes. Nu mi se pare deloc complicat. Mă veți întreba acum de ce vorbesc de o altă regulă când de fapt este exact prima regulă studiată. Pentru că această regulă spune că probabilitatea de succes a unui număr de *k* evenimente mutual exclusive reprezintă suma probabilităților de succes a fiecărui eveniment. Dacă avem *k* evenimente (în cazul nostru *k* are valoarea şase deoarece sunt şase evenimente) atunci probabilitatea de succes va fi $P_{(s)}=P_{(e1)} + P_{(e2)} + P_{(e3)} + \ldots. P_{(ek)}$

Acest lucru se verifică şi în cazul nostru, dacă ar fi să scriem formula desfăşurată:

$$P_{(s)} = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{6}{36} = \frac{1}{6}$$

Iată că v-am lămurit şi motivul pentru care această regulă se numeşte regula aditivă pentru evenimentele mutual exclusive.

Ca exercițiu, calculați probabilitatea de succes şi şansa de succes pentru a obține la table un 10 ca să scăpați de marț.

Am abordat până acum situația în care evenimentele sunt egale (cele două evenimente au acelaşi număr de rezultate – fiecare dintre cele două zaruri poate genera un număr de şase rezultate). Cum procedăm însă dacă evenimentele nu sunt egale?

Să presupunem că sunteți un jucător de poker electronic şi vă aflați în fața unui asemenea aparat care funcționează cu monede. Programatorul aparatului a implementat următorul algoritm de câştig (Dowdy, şi alții, 2004):

Careu de aşi – se câştigă 50 de monede	*probabilitate 0,001*
Careu de popi – se câştigă 30 de monede	*probabilitate 0,010*
Careu de dame – se câştigă 25 de monede	*probabilitate 0,040*
Careu de valeți – se câştigă 10 monede	*probabilitate 0,359*
Pierderea monedei	*probabilitate 0,590*

Probabilitatea de succes este practic probabilitatea de a câştiga ceva, fie valeții, damele, popii sau aşii. În acest caz probabilitatea de succes este suma probabilităților evenimentelor şi anume 0,001+0,010+0,040+0,359 rezultând 0,41.

Observăm că probabilitatea de succes, aplicând regula aditivității evenimentelor mutual exclusive, nu este altceva decât inversul probabilității de eşec. În termeni probabilistici, opusul probabilității de succes poată numele de **complement** şi se notează cu $P_{(\bar{S})}$. Întotdeauna suma dintre probabilitatea de succes şi complementul acesteia este unu. Dacă revenim la formula şansei, practic şansa de succes este raportul dintre probabilitatea de succes şi complementul acesteia.

VII.1.3 Evenimentele independente

Ați observat, din exemplul anterior cu jocul electronic, că probabilitatea de succes este destul de mare (0,41). Dacă introduceți o monedă în aparat, aveți practic o probabilitate de 41% să câştigați ceva şi 59% să pierdeți. Totuşi de ce nu câştigați? Dacă sunteți un împătimit al jocurilor de noroc, în

mod singur aţi trecut prin situaţia în care vă spuneaţi: „Am pierdut prea mult. De acum ar trebui să mai şi câştig." În realitate însă aţi risipit toţi banii fără un câştig substanţial. De ce? Răspunsul este foarte simplu. Deoarece evenimentele sunt în realitate independente. Introducerea unei noi monede în aparat nu are nici o legătură cu ceea ce aţi jucat dumneavoastră anterior. Jocurile anterioare nu influenţează cu nimic probabilitatea de câştig pe care o aveţi. În acest caz, evenimentele nu sunt mutual exclusive ci independente. Dacă veţi introduce în aparat 100 de monede, probabil că veţi câştiga de 41 de ori şi veţi pierde de 59 de ori.

Atunci când discutăm despre evenimente independente, trebuie să avem în vedere probabilitatea comună ca două sau mai multe evenimente să se întâmple simultan. Să considerăm un alt exemplu în care presupunem că avem la dispoziţie un pachet de cărţi de joc. Ştiţi foarte bine că un pachet de cărţi de joc conţine un număr de 52 de cărţi aranjate de două culori: negru şi roşu. De asemenea există, pentru fiecare culoare, un număr de 8 cărţi speciale: asul, popa, dama şi valetul. Dacă am sistematiza distribuţia cărţilor de joc, am obţine o imagine similară tabelului de mai jos (Dowdy, şi alţii, 2004):

Carte specială	**Negru**	**Roşu**	**Total**
Da	*8*	*8*	*16*
Nu	*18*	*18*	*36*
Total	*26*	*26*	*52*

În continuare, vom amesteca foarte bine cărţile din pachet şi vom încerca să găsim probabilitatea cu care putem extrage o carte specială de culoare neagră, dintr-o singură extragere. Cred că intuiţi deja răspunsul. Putem extrage oricare dintre cele 52 de cărţi din pachet. Deci totalitatea rezultatelor este de 52. Totuşi, din aceste 52 de cărţi, doar 8 cărţi sunt speciale şi negre. Prin urmare, probabilitatea de a extrage o carte specială şi de culoare neagră este de $P_{NS} = \frac{8}{52} = \frac{2}{13} = 0{,}15$. Dacă în cazul evenimentelor mutual exclusive am avut o regulă aditivă, în acest caz ne confruntăm cu o **regulă multiplica-**

tivă a evenimentelor independente. Conform acestei reguli, probabilitatea de succes a unui număr de *k* evenimente independente reprezintă produsul probabilităților de succes a fiecărui eveniment. Dacă avem *k* evenimente (în cazul nostru *k* are valoarea doi deoarece primul eveniment se referă la extragerea unei cărți de culoare neagră şi al doilea eveniment la extragerea unei cărți speciale) atunci probabilitatea de succes va fi $P_{(s)}=P_{(e1)}\ x\ P_{(e2)}\ x\ P_{(e3)}\ x\ ....\ P_{(ek)}$. În situația de față, $P_{NS} = P_N \times P_S = \frac{26}{52} \times \frac{16}{52} = \frac{1}{2} \times \frac{4}{13} = \frac{4}{26} = \frac{2}{13} = 0{,}15$. Am obținut, după cum ați constatat, exact acelaşi rezultat.

VII.1.4 Probabilitatea condiționată

Atunci când ne gândim la un african, ne vine în minte imaginea unei persoane cu tenul închis la culoare, părul negru, eventual ondulat, buzele groase. Dacă întâlnim o asemenea persoană pe stradă, putem afirma cu o probabilitate foarte mare de succes, că respectiva persoană este un african. Poate fi însă vorba însă şi despre un arab sau chiar un european. Ne-am bazat afirmația pe o serie de indicii fizice care au condiționat probabilitatea de succes. Un asemenea exemplu este un caz tipic de probabilitate condiționată. În această situație, unul sau mai multe evenimente influențează, condiționează probabilitatea de apariție a unui alt eveniment.

Să presupunem că facem un studiu pe 100 de persoane bolnave de cancer pentru a vedea în ce măsură fumatul contribuie la apariția cancerului la plămâni. După colectarea rezultatelor, le putem prezenta într-un tabel similar celui de mai jos (Dowdy, şi alții, 2004):

Cancer	**Nefumător**	**Fumător**	**Total**
Pulmonar	*5*	*20*	*25*
Alt tip de cancer	*60*	*15*	*75*
Total	*65*	*35*	*100*

Ne punem evident întrebarea, având un număr de 25 de persoane cu cancer pulmonar, care este probabilitatea ca un fumător să contacteze un can-

cer pulmonar. Răspunsul este din nou simplu dacă privim cu atenţie tabelul de mai sus. Avem un număr total de 25 de bolnavi de cancer pulmonar, din care 20 sunt fumători. În acest caz probabilitatea ca un fumător să prezinte cancer pulmonar este de $^{20}/_{25}$ = 0,80. Cam mare, nu-i aşa?

În realitate, probabilitatea condiţionată are şi ea o regulă proprie, numită **regula probabilităţii condiţionate**, care se scrie sub forma $P_{e2|e1} = \frac{P_{e1e2}}{P_{e1}}$ şi semnifică probabilitatea ca evenimentul al doilea să fie determinat de primul eveniment.

În cazul nostru, care sunt evenimentele? Lotul nostru conţine 100 de persoane, toate bolnave de cancer. Primul eveniment îl reprezintă probabilitatea ca dintre toţi bolnavii de cancer, unii să fie bolnavi de cancer la plămâni. Avem un număr de 25 de bolnavi de cancer la plămâni dintr-un număr de 100 de bolnavi de cancer. Probabilitatea va fi aşadar $^{25}/_{100}$. Al doilea eveniment este reprezentat de bolnavii de cancer la plămâni, fumători. Avem un număr de 20 de bolnavi de cancer la plămâni care fumează. Probabilitatea pentru al doilea eveniment va fi aşadar $^{20}/_{100}$. Dacă aplicăm formula de mai sus, obţinem $P_{C|F} = \frac{P_{FC}}{P_C} = \frac{\frac{20}{100}}{\frac{25}{100}} = \frac{20}{25} = 0{,}8$. Am obţinut, prin urmare, exact acelaşi rezultat.

VII.2Eroare standard şi intervale de încredere

Incursiunea anterioară în domeniul probabilităţilor, deşi puţin mai aridă faţă de cum v-aţi obişnuit, nu a fost deloc întâmplătoare şi este menită să vă familiarizeze cu demersul ştiinţific. În momentul în care utilizaţi statistica într-un studiu ştiinţific, trebuie că reţineţi că această metodă nu oferă nici pe departe concluzii certe. Orice afirmaţie pe care o facem şi care derivă din procedee statistice, are o anumită probabilitate de a fi adevărată, după cum există şi o probabilitate de a ne înşela. Acest lucru se întâmplă deoarece în demersul ştiinţific rareori lucrăm cu întreaga populaţie. De obicei studiile se

fac pe un segment din populație, numit eșantion, iar concluziile se extind la nivelul întregii populații. Devine evident faptul că indicatorii statistici obținuți prin studiul eșantionului diferă de parametrii populației generale din care acest eșantion a fost extras. Acest lucru induce anumite erori în momentul în care extindem concluziile la nivelul populației, erori numite în literatura de specialitate **erori sistematice de eșantionare**. Scopul oricărui studiu este să minimizeze aceste erori astfel încât datele obținute prin studiul eșantionului să se apropie cât mai mult de parametrii reali ai populației.

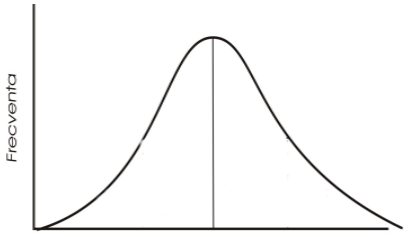


Figura 7.1 – Distribuția coeficientului de inteligență și a notelor la matematică pentru întreaga populație de elevi

Dacă vom dori, de exemplu, să studiem relația care există între coeficientul de inteligență și notele obținute la matematică de către elevii de gimnaziu din România, vom ști că cele două variabile se distribuie normal la nivelul populației, ca în figura alăturată.

Pentru a afla parametrii reali ai populației (media – μ și abaterea standard – σ) ar trebui să înregistrăm notele la matematică și rezultatele unui test de inteligență pentru toți elevii de gimnaziu din România. Un asemenea studiu va fi extrem de greu de realizat și foarte costisitor. Dacă în loc să analizăm toată populația vom studia doar 90% din populație, media obținută se va apropia foarte mult de cea a populației generale. Acest lucru devine posibil deoarece probabilitatea de a selecta subiecți

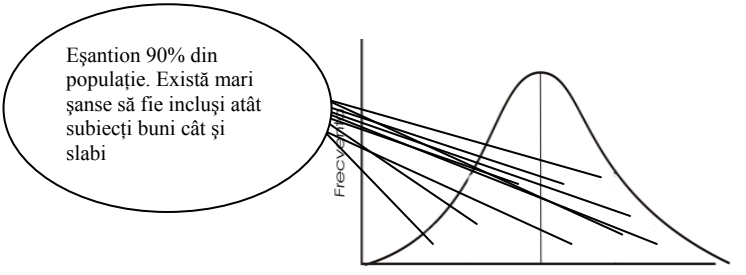


Figura 7.2 – Distribuția coeficientului de inteligență și a notelor la matematică pentru un eșantion de 90%

care vor obține scoruri egal răspândite în jurul mediei populației este foarte mare.

Însă resursele nu sunt suficiente în vederea realizării unui studiu de o asemenea amploare şi se va lucra cu un număr mai redus de subiecți. Ce se va întâmpla dacă în loc de a studia un eşantion suficient de mare vom folosi doar trei elevi? În acest caz există o probabilitate foarte mare să selectăm trei elevi cu rezultate foarte slabe sau cu rezultate foarte bune. În această situație, concluziile pe care le tragem sunt eronate, deoarece eşantionul ales nu este reprezentativ pentru populația investigată.

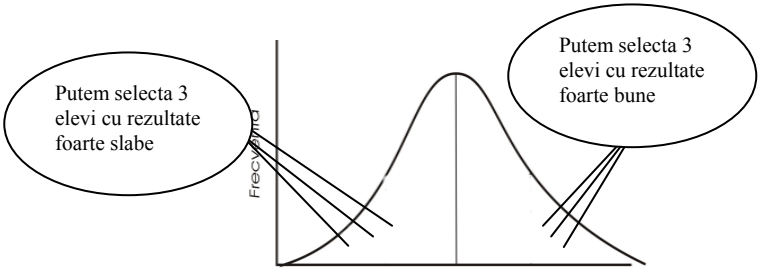


Figura 7.3 – Distribuția coeficientului de inteligență şi a notelor la matematică pentru un eşantion de 3 elevi

Pe măsură ce creştem dimensiunea eşantionului, creşte şi probabilitatea de a selecta subiecți cu scoruri deasupra şi sub media populației, fapt ce determină reducerea erorilor de eşantionare, indicatorii obținuți la nivelul eşantionului apropiindu-se din ce în ce mai mult de parametrii populației.

Într-un capitol anterior, am discutat despre teorema limitei centrale. Acest concept afirmă că dacă extragem mai multe eşantioane dintr-o populație şi calculăm media acestor eşantioane, media eşantioanelor aproximează cel mai bine media populației. În mod surprinzător, distribuția mediei eşantioanelor, în cazul unui număr mare de subiecți (peste 30 de subiecți) este o distribuție normală, indiferent de modul în care variabila se distribuie la nivelul întregii populații. Pentru a înțelege mai bine acest concept, să presupunem că avem la dispoziție un zar. După cum ştim, aruncând un zar putem obține 1, 2, 3, 4, 5 sau 6. Însumând aceste valori şi împărțind suma rezultată la şase,

rezultă media teoretică a populației de 3,5. Acum să considerăm un număr de 5 eşantioane de câte 10 aruncări cu zarul. Poate nu vă vine să credeți, însă chiar acum fac acest experiment pe care vi-l recomand şi dumneavoastră. Iată rezultatele pe care le obțin:

Eşantion 1:	1, 2, 1, 2, 6, 6, 4, 6, 4, 1	media = 3,6
Eşantion 2:	2, 2, 1, 2, 6, 5, 6, 3, 3, 4	media = 3,4
Eşantion 3:	4, 1, 2, 5, 6, 6, 3, 5, 5, 2	media = 3,9
Eşantion 4:	4, 2, 2, 2, 5, 3, 1, 4, 3, 4	media = 3,0
Eşantion 5:	5, 3, 4, 2, 6, 6, 1, 1, 4, 2	media = 3,4

Fiecare eşantion are o medie diferită de media teoretică, se abate într-un sens sau altul de la media populației 3,5. Dacă aplicăm principiile teoremei limitei centrale, vom realiza media acestor eşantioane. Adunați toate cele cinci medii şi împărțiți la cinci. Obțineți valoarea 3,46 care iată, aproximează cel mai bine media populației. În mod similar, aceste principii se aplică şi în cazul cercetărilor ştiințifice.

Media de sondaj (*m*) obținută pe baza unui eşantion reprezintă un estimator consistent, deoarece tinde spre valoarea teoretică pe măsura creşterii numărului de observații, nedeplasat, eficient, fiind estimatorul cu cea mai mică abatere de la media populației şi suficient în sensul că nici o altă estimare nu oferă informații suplimentare cu privire la parametrul estimat. (Vasilescu, 1992).

Deşi media de sondaj aproximează bine media populației, gravitând în jurul acesteia, noi nu ştim cât de bine o aproximează. Dacă veți cumpăra o pungă de cafea de 100 de grame, să nu vă aşteptați ca punga respectivă să cântărească exact 100 de grame. Veți observa pe ambalaj o specificație importantă: 100 de grame ± 5 grame. Ce înseamnă acest lucru? Înseamnă că punga dumneavoastră conține minimum 95 de grame şi maximum 105 grame de cafea. Cantitatea exactă de cafea se găseşte undeva între aceste limite, în acest interval, numit şi **interval de încredere**. Dacă vom evalua o caracteris-

tică a populației prin mai multe eşantioane, mediile fiecărui eşantion vor reprezenta estimări punctuale ale parametrului populației, estimări situate, în general, în acest interval de încredere. De aceea, putem spune că media de sondaj este un **estimator punctual** al mediei populației în timp ce intervalul de încredere reprezintă un **estimator de interval** al mediei populației. Acest indicator ne oferă un interval de valori între limitele căruia putem regăsi media reală a populației studiate.

Nu întâmplător am menționat anterior o proprietate foarte importantă a repartiției mediei eşantioanelor care tinde spre o distribuție normală, indiferent de modul în care se distribuie real variabila la nivelul populației. Cunoscând acest lucru, ne amintim că distribuția normală este complet caracterizată prin medie şi abaterea standard iar aceste concepte ne vor ajuta să determinăm intervalul de încredere.

În figura numărul 7.4, am reprezentat distribuția înălțimii bărbaților din România. Evident, media acestei caracteristici la nivelul populației este şi va rămâne necunoscută. Dacă vom extrage câteva eşantioane din populație, mediile obținute le vom putea reprezenta prin punctele de pe grafic. Media fiecărui eşantion se abate în sens pozitiv şi negativ de la media reală, însă dacă am face media acestor eşantioane ne-am apropia cel mai bine de valoarea exactă a acestui parametru.

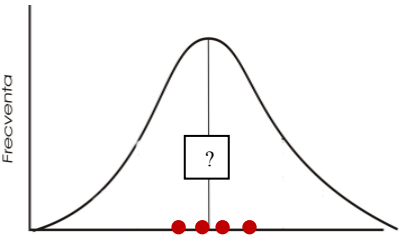


Figura 7.4 – Distribuția înălțimii bărbaților din România

Dar, după cum am spus, media eşantioanelor are proprietatea de a se distribui normal. Folosind acest element şi proprietăţile distribuţiei normale, putem calcula cu cât se abate media unui eşantion faţă de media teoretică – care rămâne în continuare necunoscută – stabilind astfel intervalul de încredere. În mod practic, mediile eşantioanelor se vor abate de la media teoretică cu un număr de… abateri standard sau, mai precis, vor dobândi, fiecare, o cotă z. Indiferent de eşantion, media acestui eşantion se poate situa între -3 şi + 3 abateri standard, pentru o probabilitate de aproximativ 99,74%. Acest interval este însă prea mare şi permite mediei de sondaj să aibă aproximativ orice valoare din domeniul de variaţie al înălţimii bărbaţilor din România. În general, trebuie să ne asigurăm, la un nivel de probabilitate de 95%, că zona din curba de distribuţie a populaţiei conţine media de sondaj, adică intervalul de încredere al mediei are o „precizie" de 95%, existând mai puţin de 5% şanse ca media populaţiei să nu fie conţinută în acest interval. În termeni de abateri standard, zona corespunzătoare unui terval de încredere de 95% se află situată între – 1,96 şi + 1,96 abateri standard.

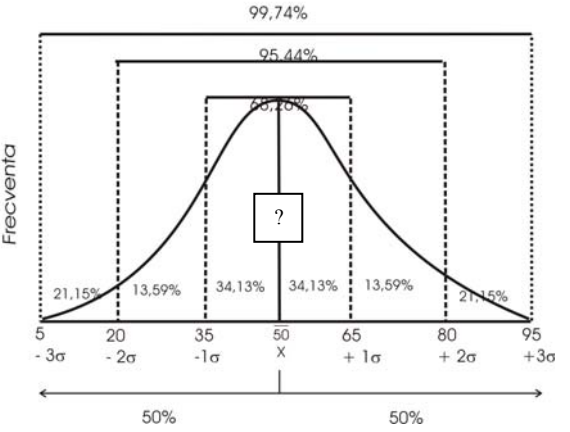


Figura 7.5 – Distribuţia normală a înălţimii bărbaţilor din România

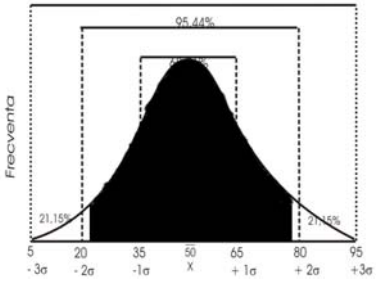


Figura 7.6 – Zona distribuţiei normale corespunzătoare unui interval de încredere de 95%

Urmând firul logic, dacă scorul z al mediei eşantionului se află situat între – 1,96 şi + 1,96 abateri standard, înseamnă că această medie aproximează suficient de bine media teoretică sau media populației – la un nivel de încredere de 95%. Dacă vă amintiți, o procedură similară am utilizat atunci când am analizat simetria şi boltirea unei distribuții.

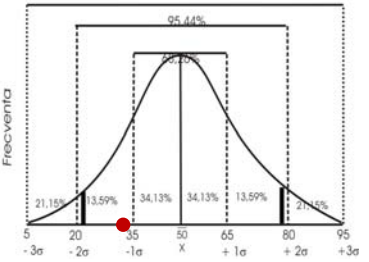


Figura 7.7 – Media eşantionului într-un interval de încredere de 95%

Abaterile mediilor de sondaj față de media teoretică seamănă foarte mult cu un alt indicator studiat anterior, ca măsură a împrăştierii, şi anume abaterea standard. De fapt, abaterea standard a mediilor eşantioanelor reprezintă un alt concept statistic care poartă numele de **eroare standard a mediei.** Eroarea standard a mediei reprezintă o măsură în care media unui eşantion deviază de la media eşantioanelor, având o importanță deosebită în stabilirea reală a intervalului de încredere şi a măsurii în care media unui eşantion aproximează media reală a populației.

Metoda tradițională prin care am putea estima eroarea standard a mediei ar fi aceea de a extrage mai multe eşantioane din populație, de a calcula media eşantioanelor şi apoi abaterea standard a acestora după unul dintre procedeele descrise în capitolele anterioare. Totuşi, acest lucru nu se întâmplă, deoarece, de cele mai multe ori, în cercetarea ştiințifică lucrăm cu un singur eşantion. Din fericire, eroarea standard a mediei este în legătură cu mărimea eşantionului – aşa cum am arătat anterior – şi poate fi calculată cunoscând doar mărimea eşantionului şi abaterea standard a acestuia. Cu cât eşantionul este mai mare, cu atât eroarea standard este mai mică şi invers. O bună aproximare a erorii standard a mediei este dată de raportul dintre abaterea standard a rezultatelor obținute de către subiecții din eşantion şi rădăcina pătrată din volumul eşantionului.

Eroarea standard a mediei poate fi aşadar calculată prin intermediul formulei:

$$S_m = \frac{\sigma}{\sqrt{n}} \qquad \text{(formula 7.1)}$$

unde S_m reprezintă eroarea standard a mediei, σ reprezintă abaterea standard a scorurilor la variabila analizată şi n reprezintă volumul eşantionului.

Vom considera valorile ce reprezintă înălțimea subiecților exprimată în centimetri: *182, 175, 174, 189, 177, 177, 180, 173, 188, 171, 184, 192, 180, 172, 177, 193, 184, 192, 172, 177, 174, 171, 172, 179, 185, 179, 173, 172, 179, 188.* Ştim că am măsurat un eşantion de 30 de bărbaţi şi am obținut scorurile anterioare, distribuția având media de 183,2 centimetri şi abaterea standard de 5,23 centimetri.

Dacă aplicăm acum formula, vom obține eroarea standard a mediei de $S_m = \frac{5{,}23}{\sqrt{30}} = \frac{5{,}23}{5{,}47} = 0{,}95$. Într-adevăr, în acest caz eroarea standard a mediei este destul de mică, media aproximând bine media generală a populației.

Cunoscând eroarea standard a mediei, putem acum estima limitele între care găsim media populației (μ) la un interval de încredere de 95%. După cum ştim, în termeni de abateri standard, pe o distribuție normală, intervalul de încredere corespunzător nivelului de 95% este situat între +/- 1,96 abateri standard. Particularizând, intervalul de încredere al mediei teoretice poate fi calculat după formula:

$$\mu = m \mp 1{,}96\frac{\sigma}{\sqrt{n}} = m \mp 1{,}96 \times S_m \qquad \text{(formula 7.2)}$$

În cazul nostru, media înălțimii celor 30 de subiecți este de 183,5 centimetri iar eroarea standard a mediei este de 0,95 centimetri. Limita inferioară a intervalului de încredere va fi de 183,5 – 1,96x0,95 = 181,67 centimetri iar limita superioară a intervalului de încredere va fi de 185,36 centi-

metri. Cu alte cuvinte, între 181,67 centimetri şi 185,36 centimetri se va afla media de înălțime a populației masculine din România.

VII.3Metoda ştiinţifică în ştiinţele socio-umane

Având în vedere informațiile prezentate până acum, ar fi momentul să discutăm despre metoda ştiinţifică pe care o utilizăm ori de câte ori suntem în fața desfăşurării unui studiu serios şi valid în ştiinţele socio-umane. Abordez această problemă pentru că, am văzut de multe ori, în rândul cercetătorilor sau al cadrelor universitare, opinii cel puțin paradoxale. Ce spuneți, de exemplu, de o replică de genul: „Aplicați instrumentele, vedem ce iese şi după aceea formulăm ipotezele". Nu ştiu dacă v-ați confruntat sau nu cu o asemenea atitudine. Eu m-am confruntat pe parcursul multor lucrări de licență şi chiar teze de doctorat şi am rămas stupefiat de lipsa elementară de cunoştințe în ceea ce priveşte metoda ştiinţifică. Este la fel ca şi cum am spune: „Construim acoperişul, vedem cum arată şi după aceea facem fundația şi pereții". Un asemenea demers este unul diletantist, amatoricesc, care nu are nici o legătură cu cercetarea ştiinţifică. Demersul ştiinţific presupune, de fiecare dată, un proces strict etapizat, care se raportează la următoarele aspecte (Dowdy, şi alții, 2004):

1. Enunțarea problemei;
2. Formularea ipotezelor;
3. Proiectarea cercetării;
4. Efectuarea observațiilor;
5. Interpretarea datelor;
6. Formularea concluziilor.

Deşi nu intenționez abordarea în detaliu a tuturor etapelor, aş dori să menționez faptul că, în principal, metoda statistică intervine în etapa a cincia, etapa interpretării datelor. Deoarece o interpretare a datelor presupune însă şi

o colectare corectă a lor, cunoştinţe de statistică sunt necesare atât în etapa a doua – formularea ipotezelor – cât şi în etapa proiectării experimentului sau a sondajului. Numai în acest mod ne putem asigura de prezenţa unui set valid de date care va face posibilă interpretarea acestora şi transferul către etapa a şasea – etapa formulării concluziilor. Iată motivul pentru care fiecare dintre aceste şase etape vor fi analizate separat.

VII.3.1 Enunţarea problemei

Enunţarea problemei vi se pare, probabil, cea mai simplă etapă a demersului ştiinţific. Trebuie să vă informez, de la început, că acest stadiu nu este deloc simplu şi necesită un mare volum de cunoştinţe, o bună intuiţie ştiinţifică, un volum impresionant de observaţii şi, de asemenea, un potenţial creativ. Pot, spre exemplu, să enunţ problema călătoriei în timp. Demersul ştiinţific va eşua din start, deoarece, la ora actuală, problema este insolvabilă. O problemă apare în urma unui mare număr de observaţii. Pot constata, de exemplu, că un element major al ineficienţei instituţionale îl reprezintă incapacitatea de comunicare la diferitele structuri ierarhice din cadrul instituţiei. Această problemă, pe care mi-o propun spre soluţionare, rezultă în urma unui număr semnificativ de observaţii referitoare la analiza dinamicii şi a climatului organizaţional la nivelul multiplelor instituţii. În acelaşi timp, trebuie să ne asigurăm de faptul că problema nu are încă o soluţie pertinentă. Degeaba observ că stând la umbra unui măr, îmi cade un fruct în cap. Poate voi fi capabil să mă întreb de ce cade, să îmi pun problema şi apoi să găsesc explicaţia. Însă această problemă a fost rezolvată acum mult timp de către Newton. Referitor la acest aspect, investigarea soluţiilor pertinente este de domeniul metaanalizei, demers ce ar trebui să preceadă etapa enunţării problemei şi pe care îl vom aborda în al treilea volum al prezentei lucrări.

Legăturile omului de ştiinţă cu domeniul investigat sunt, din nou, de mare importanţă. Am depăşit epoca renascentistă în care un om putea fi foarte bun în chimie, fizică, anatomie, arte, matematici şi aşa mai departe. Volu-

mul de cunoştinţe acumulat de omenire este atât de mare, încât chiar într-un subdomeniu limitat, fiinţa umană este depăşită. Vorbind de psihologie – ca să nu luăm în discuţie domeniul supraordonat al ştiinţelor socio-umane – numai un diletant poate pretinde că e… psiholog. Este ca şi cum aş spune că un medic este pur şi simplu medic. Evident că e medic, însă în ce? În ortopedie, pediatrie, neurologie, oftalmologie şi aşa mai departe. La fel, un psiholog poate fi psihoterapeut, experimentalist, consilier şcolar, specialist în organizaţii, în psihologia socială etc. Mie, de exemplu, îmi place psihologia experimentală şi în special metoda statistică. Totuşi, la momentul în care scriu aceste rânduri, consider că ştiu cel mult 15-20% din metoda statistică şi mă întreb dacă îmi va ajunge viaţa să deţin un procent de 50-60%, deşi studiul acestei metode face parte din activitatea mea zilnică. În aceste condiţii, cum aş putea oare să formulez o problemă din domeniul psihoterapiei, ca să nu mai vorbim de domeniul chimiei sau al astronomiei. Poate pare deplasat, însă sunt adeptul strictei specializări şi al lucrului în echipe multidisciplinare. Dacă sunt psihoterapeut, desigur, am anumite cunoştinţe de statistică psihologică, însă nu mă pot considera un expert în această problemă. Este de preferat lucrul într-o echipă cu un specialist în psihologie experimentală, care stăpâneşte la un nivel perfect satisfăcător metoda statistică, în momentul în care apare necesitatea unui studiu ştiinţific ce presupune analiza datelor, în locul unui efort pe termen scurt de abordare pe cont propriu al unui domeniu în care competenţele sunt mediocre.

Dincolo de aceste menţiuni, enunţarea problemei presupune formularea acesteia în scris, în termeni clari şi expliciţi. În urma formulării problemei, rezultă obiectivele cercetării, obiective exprimate în termeni de obiective generale şi specifice.

Obiectivele generale ghidează cercetarea în ansamblul ei. Un studiu ştiinţific are unul, maxim două obiective generale. Nu ne putem concentra eforturile pe mai multe planuri – ori planurile sunt ghidate exact de modul de formulare al obiectivelor generale.

Obiectivele generale ale unei cercetări sunt formulate în urma unui cumul de observații şi, eventual, în urma unui studiu metaanalitic. În ultima situație, este necesară prezentarea rezultatelor studiului metaanalitic care a ghidat formularea obiectivului general şi prin care se demonstrează lipsa de abordare sau abordarea nepertinetă a problemei studiate. Spre exemplu, constatăm că pe măsură ce creşte numărul de kilometri parcurşi cu un automobil, creşte şi riscul de apariție al unui accident şi ne propunem să studiem ştiințific această problemă. În mod normal, ar trebui să începem prin a analiza cercetările existente în acest domeniu şi să desfăşurăm un studiu metaanalitic. În urma acestui studiu, putem preciza obiectivul general al cercetării, de forma: *„Obiectivul general al cercetării îl reprezintă stabilirea relației existente între numărul de kilometri parcurşi şi probabilitatea de apariție a unui accident rutier”*. Observăm că enunțarea problemei ce urmează a fi studiată s-a făcut extrem de clar şi explicit, astfel încât se pot defini termenii şi condițiile proiectării unui studiu experimental sau a unei cercetări bazate pe sondaj.

Obiectivele specifice reprezintă aspecte detaliate ale investigației ştiințifice, derivă din obiectivul general şi creează baza formulării ipotezelor de cercetare (alternative). În funcție de dimensiunile cercetării şi de resursele pe care le are la dispoziție cercetătorul, se pot formula unul, două, trei… douăzeci de obiective specifice, aflate în legătură cu obiectivul general.

Concluzionând, enunțarea problemei reprezintă prima etapă a demersului ştiințific. Ea este însoțită de un studiu al cercetărilor în domeniu şi de prezentarea clară şi precisă, în termeni cuantificabili, a obiectivelor generale şi specifice.

VII.3.2 Formularea ipotezelor

A doua etapă a demersului ştiințific este reprezentată de formularea ipotezelor cercetării. *O ipoteză reprezintă o prezumție clară, explicită şi verificabilă referitoare la relațiile sau diferențele existente între două sau mai multe variabile.* Plecând de la definiția menționată mai sus, rezultă faptul că o

ipoteză de cercetare este legată, evident, de obiectivele cercetării. Deoarece o ipoteză urmează să fie testată statistic, formularea acesteia trebuie făcută în termeni de claritate maximă, fără ambiguităţi, demersul testării ipotezelor fiind unul logic şi demonstrabil prin procedee statistice specifice.

Deoarece formularea ipotezelor şi înţelegerea logicii testării acestora reprezintă un element vital în cercetarea ştiinţifică, vom analiza pe larg modalitatea prin care sunt elaborate şi verificate ipotezele.

Ipotezele sunt formulate atât în cazul studiilor care urmăresc stabilirea unor relaţii dintre variabile (numite şi studii corelaţionale) cât şi pentru cercetările ce vizează existenţa unor diferenţe dintre variabile (studii factoriale). O primă distincţie se poate face între ipoteze unilaterale (unidirecţionale) şi ipoteze bilaterale (bidirecţionale) (Clocotici, şi alţii, 2000).

- **Ipotezele unilaterale** se utilizează în momentul în care avem o idee de sensul, direcţia în care evoluează variabilele. Aceste ipoteze sunt mai precise şi permit dezvoltarea unor studii pertinente, existând şanse mai mari să fie susţinute de analiza datelor. O ipoteză unilaterală se recunoaşte după modul de formulare, deoarece în enunţul acesteia regăsim direcţia de evoluţie a variabilelor. De exemplu, dacă desfăşurăm o cercetare care are ca scop analiza legăturii dintre anxietate şi depresie, am putea formula o ipoteză unidirecţională de tipul: *„există o legătură pozitivă între nivelul anxietăţii şi nivelul depresiei”*. În această situaţie, ipoteza va fi susţinută numai în cazul în care legătura dintre anxietate şi depresie este pozitivă, adică în situaţia în care subiecţii cu anxietate mare manifestă şi puternice simptome depresive sau subiecţii cu anxietate mică nu prezintă simptome depresive. Ipoteza nu se susţine dacă, de exemplu, subiecţii cu anxietate mică manifestă puternice simptome depresive (cazul unei corelaţii negative). Deşi mult mai precisă,

există în permanență riscul să nu putem susține ipoteza în condițiile în care sensul formulat nu este reprezentat de datele cercetării, chiar dacă în urma analizei pot rezulta şi alte sensuri specifice;

- **Ipotezele bilaterale** nu impun direcția de evoluție a variabilelor. Deşi mai puțin precise în comparație cu cele unilaterale, ipotezele bilaterale oferă o mai mare libertate cercetătorului. O formulare de genul *„există o legătură între nivelul de anxietate şi predominanța simptomelor depresive"* poate fi susținută atât în cazul în care corelația este pozitivă cât şi în situația unei corelații negative. Ipoteza nu este susținută doar în condițiile în care, cu o mare probabilitate, se demonstrează inexistența unei asemenea relații.

Chiar dacă vi se poate părea puțin ciudat, testarea unei ipoteze nu se face prin formularea menționată mai sus – denumită şi **ipoteză alternativă** – ci printr-un fel de „invers" al acesteia pe care îl numim în statistică **ipoteza nulă** sau **ipoteza statistică.** Testarea unei ipoteze înseamnă, de fapt, testarea ipotezei nule în vederea respingerii sau nerespingerii acesteia.

VII.3.2.1 Ipoteza nulă şi ipoteza alternativă

Înainte de a aborta testarea ipotezelor, ar trebui să înțelegem foarte clar ce înseamnă ipoteza nulă şi ipoteza alternativă.

Reluând exemplul de mai sus, ipoteza alternativă se notează, de obicei, cu H_1 şi poate fi enunțată astfel:

„H_1: Presupunem că există o legătură între nivelul de anxietate şi nivelul de depresie."

Din punct de vedere statistic, nu putem verifica direct această ipoteză. Singura modalitate prin care putem sprijini ipoteza alternativă este să enun-

ţăm şi apoi să testăm ipoteza nulă. De obicei, ipoteza nulă – notată cu H_0 – este o negare logică a ipotezei alternative şi poate fi formulată astfel:

„H_0: Nu există nici o legătură între nivelul de anxietate şi nivelul de depresie."

Într-o cercetare, există o singură ipoteză nulă şi una sau mai multe ipoteze alternative. Ipoteza nulă este cea care orientează planul (designul) cercetării. Dacă în urma analizei datelor respingem ipoteza nulă, înseamnă că datele noastre susţin una dintre ipotezele alternative. Dacă ipoteza nulă nu este respinsă, datele nu sprijină ipoteza alternativă.

Acesta este singura modalitate prin care se poate testa o ipoteză. Reţineţi că nu putem confirma sau infirma ipoteza alternativă şi nici ipoteza nulă. ***Singurul lucru pe care îl putem face este că respingem sau să nu respingem ipoteza de nul.*** Nu vă revoltaţi. Acesta este adevărul, chiar dacă nu vă place. Pentru ca acest lucru să vă devină clar, vom explica puţin, prin exemplu, logica testării ipotezei de mai sus. Cum aţi proceda pentru a studia legătura dintre anxietate şi depresie? Vom lua cazul cel mai simplu, în care dispunem de două chestionare, unul care măsoară anxietatea şi un al doilea care măsoară depresia. Administrăm apoi cele două chestionare unui eşantion de, să zicem, 30 de subiecţi. După colectarea datelor, plecăm de la ideea că nu există nici o legătură între nivelul de anxietate şi nivelul depresiei.

Nu trebuie să uităm faptul că scopul acestei cercetări îl reprezintă extragerea unei concluzii valabilă la nivelul întregii populaţii pe baza studiului unui eşantion. Chiar dacă observăm o relaţie între cele două variabile, întotdeauna există o probabilitate – mai mare sau mai mică – ca rezultatul să fie obţinut pe baza unei erori de eşantionare. Ipoteza nulă ne indică, de fapt, probabilitatea de a obţine aceste date în condiţiile în care ar fi adevărată, adică în condiţiile în care datele se obţin în urma unei erori de eşantionare. Practic, în momentul în care calculăm relaţia dintre cele două variabile, calculăm probabilitatea de obţinere a acestei relaţii ca rezultat al unei erori de eşantionare.

Oricând există posibilitatea ca relația constatată la nivelul eşantionului să nu se regăsească într-adevăr la nivelul populației, deoarece datele colectate la nivelul eşantionului sunt rezultate în urma unei erori de eşantionare.

În exemplul nostru, să presupunem că găsim o relație între anxietate şi depresie, în sensul că persoanele anxioase prezintă şi simptome depresive. Am lucrat totuşi pe un eşantion şi nu cu întreaga populație. Înainte de a face această afirmație valabilă pentru întreaga populație, ne putem pune problema care ar fi probabilitatea ca relația pe care am descoperit-o să nu existe de fapt (ipoteza nulă), adică relația să fie rezultatul unor erori de eşantionare – este, evident, o probabilitate condiționată. Dacă această probabilitate este suficient de mică, să spunem 5%, atunci putem concluziona că relația descoperită nu este rodul unor erori de eşantionare ci o putem regăsi la nivelul populației. Acest procent de 5% nu înseamnă altceva decât că, la doar 5 subiecți dintr-o sută de subiecți relația s-ar putea să nu se verifice. În acest caz, probabilitatea ca ipoteza nulă să fie adevărată este foarte mică şi o putem respinge. Respingând ipoteza nulă, putem sprijini ipoteza alternativă. Nu am spus că o confirmăm sau că o acceptăm. În realitate ipoteza alternativă nu poate fi confirmată sau acceptată, deoarece în permanență apare excepția (cele 5 persoane dintr-o sută) pentru care ipoteza alternativă nu este adevărată.

Sperăm că în baza exemplului de mai sus ați reuşit clarificarea modului în care are loc testarea ipotezelor într-o cercetare ştiințifică. Dacă ar fi să concluzionăm, am putea spune că logica testării ipotezelor cuprinde patru etape (Dancey, şi alții, 2002):

- Formularea ipotezei;
- Măsurarea variabilelor implicate şi stabilirea relațiilor dintre variabile;
- Calcularea probabilității de obținere a acestor relații în condițiile în care relațiile de fapt nu există la nivelul populației;

- Dacă probabilitatea calculată în etapa anterioară este destul de mică, atunci există o probabilitate foarte mare ca relaţia descoperită să se regăsească la nivelul populaţiei.

O ultimă menţiune referitoare la ipoteza nulă. În general, cercetătorii consideră ipoteza nulă o ipoteză a absenţei. (Sava, 2004). Această abordare este o abordare eronată. Ipoteza nulă nu înseamnă lipsa unei corelaţii sau lipsa unei diferenţe dintre medii ci probabilitatea ca diferenţa sau corelaţia să fie obţinută ca rezultat al unei erori de eşantionare. O legătură între două variabile, chiar foarte scăzută, există totuşi, doar că probabilitatea ca această relaţie să fie rezultatul unor erori de eşantionare este foarte mare şi nu trebuie ignorată. Prin urmare, ipoteza nulă nu afirmă de fapt că nu există o legătură ci trebuie privită mai degrabă ca o ipoteza ce urmează să fie anulată şi nu ca ipoteza absenţei.

VII.3.2.2 Semnificaţie şi eroare în testarea ipotezelor

Am afirmat că putem respinge ipoteza nulă (ipoteza conform căreia relaţia obţinută este determinată de erorile de eşantionare) dacă probabilitatea să fie adevărată este foarte mică. Bine, mă veţi întreba, dar cât de mică? Care este pragul la care putem respinge ipoteza nulă? În cercetarea ştiinţifică lucrăm cu două asemenea praguri de semnificaţie. Primul prag a fost postulat de Fisher şi reprezintă probabilitatea de 0,05 (5% sau 1 caz din 20 de cazuri) iar al doilea prag de semnificaţie reprezintă probabilitatea de 0,01 (1% sau 1 caz din 100 de cazuri).

Pragul de semnificaţie reprezintă probabilitatea (de 5% sau 1%) ca ipoteza nulă să nu fie respinsă sau, cu alte cuvinte, la 1 din 20 de cazuri sau la 1 din 100 de cazuri relaţia să nu existe ori să fie determinată de erorile de eşantionare.

Pragul de semnificaţie se notează, în literatura de specialitate, cu p sau cu α şi permite respingerea ipotezei nule în condiţiile unor valori strict mai

mici decât *p* (atunci când *p<0,05*). De asemenea, un indicator statistic poate fi **semnificativ** sau **nesemnificativ** în funcție de raportarea la acest prag.

Presupunând că ipoteza nulă este adevărată, dacă probabilitatea ca relația să fie obținută prin intermediul erorilor de sondaj este mai mică de 0,05 (5%) atunci se poate respinge ipoteza nulă iar indicatorul statistic este semnificativ. În exemplul nostru, putem respinge ipoteza nulă şi putem afirma că *există o legătură semnificativă între nivelul anxietății şi nivelul depresiei.* În aceleaşi condiții, dacă probabilitatea ca relația să fie obținută prin intermediul erorilor de sondaj este mai mare de 0,05, indicatorul nu este semnificativ iar ipoteza nulă nu poate fi respinsă. Vom putea atunci afirma că *nu există nici o legătură semnificativă între nivelul anxietății şi nivelul depresiei.*

În orice lucrare ştiințifică, alături de valoarea indicatorului statistic se raportează şi valoarea pragului de semnificație. Există două modalități prin care puteți raporta pragul de semnificație:

- prin raportarea inegalității, de forma p<0,05 sau p<0,01. Această modalitate, deşi încă practicată, nu este extrem de agreată în comunitatea ştiințifică deoarece permite o raportare grosieră a probabilității de respingere a ipotezei de nul. Un p<0,05 poate însemna atât o probabilitate de 0,49 dar şi o probabilitate de 0,11;

- prin raportarea exactă a pragului de semnificație, de forma p=0,38. Această metodă este preferată, deoarece permite o evaluare exactă a probabilității de respingere a ipotezei de nul. Unele programe statistice, printre care şi SPSS, raportează uneori pragul de semnificație de forma .000, adică 0,000. Este o greşeală să precizați un p=0,000, acest lucru însemnând o probabilitate nulă de obținere a relației prin eroarea de eşantionare, ceea ce nu poate fi posibil. Deoarece SPSS rotunjeşte la trei zecimale, probabil că pragul real de semnificație este o va-

loare de genul 0,0000001457. În acest caz, se preferă prima modalitate de raportare, de genul $p<0,01$.

Atunci când discutăm de semnificație, trebuie să avem în vedere faptul că ne referim la o **semnificație statistică** şi nu la semnificația psihologică, socială sau economică. Afirmația conform căreia există o corelație semnificativă între nivelul de anxietate şi nivelul depresiei la un prag de semnificație mai mic de 0,05 nu înseamnă altceva decât că, la nivelul populației, există o probabilitate mai mică de 5% ca să nu fie nici o relație între anxietate şi depresie. Atât. Semnificația statistică nu implică o interpretare psihologică a acestei legături.

Valoarea pragului de semnificație reprezintă o probabilitate condiționată, probabilitatea de apariție a unui eveniment în condițiile în care ipoteza nulă este adevărată. Această valoare nu indică probabilitatea de apariție a evenimentului la nivelul populației. Un prag de semnificație mai mic de 0,05 ne spune doar că sunt mai puțin de 5% şanse ca relația dintre anxietate şi depresie să fie rezultatul unei erori de eşantionare. Acest lucru însă nu înseamnă că avem peste 95% şanse să regăsim relația la nivelul populației. După cum am mai afirmat, testarea ipotezelor presupune doar testarea ipotezei nule nu şi inferențe referitoare la ipoteza alternativă.

Chiar dacă pragul de semnificație creează condițiile respingerii ipotezei nule, totuşi nu vom putea niciodată afirma cu certitudine că susținem ipoteza alternativă. În cercetarea ştiințifică, deşi condițiile de semnificație statistică pot fi îndeplinite, este posibilă apariția a două tipuri de erori: putem respinge ipoteza nulă iar în realitate ea să nu îndeplinească condițiile de respingere la nivelul populației sau putem să nu respingem ipoteza nulă în situația în care ar trebui respinsă.

Când credeați şi dumneavoastră că ați scăpat de lucrurile astea ce presupun logica probabilităților, iată un nou concept care are darul să vă amețească. Personal mă mir că nu ați trecut încă la lectura unui volum de poveşti.

Şi dacă tot veni vorba de poveşti, să vă relatez o poveste pe care mi-a spus-o pe vremuri Ovidiu Lungu şi care vă va lămuri cu privire la natura acestor erori.

A fost odată un împărat care avea obiceiul să poarte multe războaie. Înainte de a merge la război, împăratul nostru chema vrăjitorul curții şi îl punea să-i prezică soarta bătăliei. De fiecare dată, bietul vrăjitor era în mare impas deoarece dacă greşea previziunea risca să-şi piardă capul cum o pățiseră mulți alții înaintea lui. Totuşi, vrăjitorul avea mulți ani de când îşi păstra capul pe umeri iar predicțiile acestuia, uneori se dovedeau a fi adevărate. Întrebarea mea este cum proceda? În primul rând, vrăjitor fiind, el citise această carte şi celelalte două care urmează să apară. În al doilea rând, avea ceva noțiuni legate de probabilități şi verificarea ipotezelor şi înaintea bătăliei îşi formula ipotezele, apoi construia un tabel similar tabelului de mai jos:

H_1: Prevăd că măria sa va câştiga bătălia

H_0: Prevăd că măria sa nu va câştiga bătălia

		Ce se întâmplă în urma bătăliei	
		Câştigă	**Pierde**
Ce a prevăzut că se întâmplă	**Câştigă**	OK p= 1-α Caracteristica testului	Eroare tip I p=α Pragul de semnificație
	Pierde	Eroare tip II p=β	OK p=1-β Puterea testului

Să analizăm acum tabelul de mai sus. Observăm că există un număr de patru situații:

- Vrăjitorul a prezis că împăratul va câştiga bătălia şi împăratul a câştigat-o. Aceasta este situația descrisă până acum. Vrăjitorul respinge ipoteza nulă şi susține ipoteza alternativă deoarece probabilitatea de a nu câştiga bătălia era foarte mică. Celălalt împărat avea probabil un număr mic de oameni şi o tehnică de

luptă net inferioară. Vrăjitorul a obținut o nouă avere şi respectul împăratului;

- Vrăjitorul a prezis câştigarea bătăliei şi împăratul a pierdut-o. S-a comis astfel o eroare de tip I, respingându-se ipoteza nulă când de fapt ar fi trebuit acceptată. Pragul de semnificație a fost probabil apropiat de limită (0,05) însă vrăjitorul a riscat. Şi-a pierdut şi averea şi capul. Este cea mai gravă eroare pe care o poate face. Probabil că s-a bazat pe faptul că celălalt împărat are puțini oameni şi o tehnică de luptă inferioară însă l-au dezinformat spionii şi a subapreciat fanatismul ostaşilor;
- Vrăjitorul a prezis pierderea bătăliei şi împăratul a câştigat-o. S-a comis acum o eroare de tip II, nerespingând ipoteza nulă în condițiile în care ar fi trebuit respinsă. În bucuria victoriei, împăratul s-ar putea să-i cruțe capul vrăjitorului însă va pierde respectul şi o parte din avere. Vrăjitorul nu a vrut să rişte deoarece cunoştea fanatismul luptătorilor celuilalt împărat;
- Vrăjitorul a prezis pierderea bătăliei şi împăratul a pierdut-o. Din fericire şi-a păstrat şi capul şi averea deoarece a avut înțelepciunea că calculeze puterea unui test statistic, adică probabilitatea de a respinge ipoteza nulă atunci când ea este falsă.

După cum ați putut observa, eroarea de tip I este situația în care un cercetător respinge ipoteza nulă în condițiile în care aceasta nu ar fi trebuit respinsă iar eroarea de tip doi este situația în care nu se respinge ipoteza nulă deşi ar trebui respinsă. Cele două tipuri de erori se află într-un raport invers proporțional. Pe măsură ce cresc şansele de apariție a erorii de tip I, scad şansele de apariție a erorii de tip II şi invers.

Din nefericire, cele mai multe cercetări riscă apariția unei erori de tip I, care am văzut că este şi cea mai gravă, neglijând posibilitatea de apariție a erorii de tip II.

Eroarea de tip I corespunde pragului de semnificație de 0,05 (5%). Avem 5% şanse să respingem ipoteza nulă în condițiile în care s-ar dovedi adevărată. O probabilitate cam mare atunci când e în joc capul vrăjitorului.

Eroarea de tip doi corespunde unui prag de semnificație de 0,01 (1%). De data aceasta avem doar 1% să respingem ipoteza nulă în condițiile în care s-ar dovedi adevărată, însă mult mai multe şanse să nu o respingem când ar trebui respinsă şi astfel să comitem o eroare de tip II. Oricum, consecințele ar fi mult mai blânde.

Puteți acum să-mi spuneți cum de şi-a păstrat vrăjitorul capul pe umeri atâția ani? Intuiți deja că a lucrat la un prag de semnificație de 0,01, în care creşte riscul apariției erorii de tip II, însă scade probabilitatea unei erori de tip I.

Rolul acestor erori este deosebit de important în cercetare. Să presupunem că o firmă de medicamente descoperă un nou produs care opreşte evoluția cancerului însă are efecte secundare în ceea ce priveşte viața pacienților deoarece generează o serie de produse toxice. Problema care se pune este la ce nivel de risc pot fi acceptate efectele toxice în raport cu beneficiile pe care le poate aduce medicamentul. În urma unui studiu, pot exista două posibilități:

- Prima posibilitate este cea corespunzătoare unei erori de tip I. Medicamentul să fie pus pe piață în condițiile în care pragul de semnificație este de 0,05. S-a respins ipoteza nulă conform căreia medicamentul este toxic şi s-a optat în favoarea beneficiilor pe care le aduce. La un asemenea prag de semnificație s-ar

putea ca la nivelul populației tratamentul să provoace mai multe decese decât vindecări;

- A doua posibilitate este cea corespunzătoare unei erori de tip II. Medicamentul să nu fie pus pe piață deoarece pragul de semnificație a fost mai mare de 0,01. Deşi medicamentul poate opri evoluția cancerului, totuşi riscul efectelor toxice este mult prea mare pentru a fi acceptat. Oricum, este o situație mult mai convenabilă în raport cu primul caz.

Cunoscând aceste tipuri de erori, cum le putem totuşi evita? Modalitatea cea mai bună de evitare a erorilor este replicarea cercetării prin proiectarea unui alt studiu. Dacă în urma unui alt studiu rezultatele se păstrează, nivelul de încredere în concluziile extrase va creşte. Dacă observăm o singură dată un fenomen, există un mare risc ca această observație să se datoreze întâmplării. Dacă îl observăm de două, trei, patru ori, putem afirma cu mai mare certitudine că reprezintă un fenomen real, reproductibil şi generalizabil.

VII.3.2.3 Ipoteze unilaterale şi bilaterale

Importanța acestei clasificări a ipotezelor de cercetare implică precizări suplimentare, în afara celor menționate în secțiunea anterioară. Am stabilit deja că putem formula unidirecțional o ipoteză, atunci când precizăm direcția exactă a relațiilor dintre două variabile (*„Există o corelație pozitivă puternică între anxietate şi depresie”*) sau bidirecțional, atunci când nu putem anticipa exact natura relațiilor dintre variabile (*„Există o corelație între anxietate şi depresie”*). În literatura de specialitate prima situație poartă numele de **„one-tailed hypothesis”** iar a doua situație **„two-tailed hypothesis”** sau, într-o variantă românească aproximativă „ipoteze la un capăt la distribuției - unilaterale” şi „ipoteze la ambele capete ale distribuției - bilaterale”. Această denumire o întâlnim şi în cadrul programului SPSS şi nu ar fi rău să aflăm logica ce stă în spatele acestor nume.

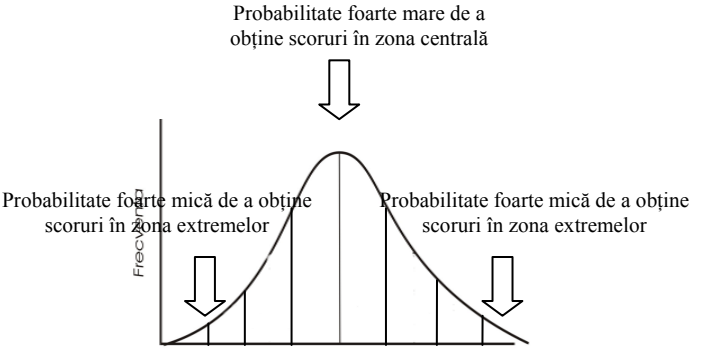

Probabilitate foarte mare de a obține scoruri în zona centrală
Probabilitate foarte mică de a obține scoruri în zona extremelor
Probabilitate foarte mică de a obține scoruri în zona extremelor
Frecvența

Figura 7.8 – Probabilitatea de a obține scoruri în zona centrală a distribuției în comparație cu extremele

Știm deja că distribuția normală prezintă două capete, două extreme care se apropie de frecvența zero pe care nu o ating însă niciodată și am studiat deja probabilitatea de a obține scoruri la capetele distribuției în comparație cu probabilitatea de a obține scoruri în zona centrală. Care ar fi probabilitatea de a întâlni o persoană cu înălțimea de 178 de centimetri? Ne aducem aminte că această valoare reprezintă o înălțime comună, cu probabilitate de apariție mare, o persoană cu această înălțime regăsindu-se undeva în zona din mijloc a distribuției. O persoană cu înălțimea de 250 de centimetri are o probabilitate de apariție foarte mică, situată undeva la capătul din dreapta a distribuției în timp ce o persoană cu înălțimea de 120 de centimetri are de asemenea o probabilitate mică de apariție, undeva la capătul din stânga al distribuției. La fel procedăm și în cazul în care discutăm de o altă variabilă, să spunem coeficientul de inteligență. Un IQ de 110 este unul comun, pe care îl putem întâlni în zona de probabilitate mare. Un IQ de 180 este unul foarte mare, cu o probabilitate de apariție mică, similară înălțimii de 250 de centimetri. La fel, un IQ de 60 are o probabilitate mică de apariție și se situează spre extrema stângă a distribuției.

Haideți să ne punem acum problema relației dintre cele două variabile. Ce se poate întâmpla cu coeficientul de inteligență pe măsura variației înălțimii subiecților? Pot exista mai multe variante posibile:

- Pe măsură ce crește înălțimea, crește și coeficientul de inteligență. Persoanele cu înălțime mare vor avea un coeficient de

ță scăzut. Spunem în acest caz că creşterea scorurilor la o variabilă determină creşterea scorurilor la cea de-a doua variabilă;

- Pe măsură ce creşte înălțimea, scade coeficientul de inteligență. Persoanele cu înălțime mare au un coeficient de inteligență scăzut în timp ce persoanele scunde au un coeficient de inteligență ridicat. Spunem în acest caz că creşterea scorurilor la o variabilă determină micşorarea scorurilor la a doua variabilă.
- Nu există nici o relație dintre cele două variabile. Întâlnim persoane înalte cu coeficienți de inteligență mari şi mici, la fel şi în cazul persoanelor scunde sau cu o înălțime medie. În acest caz spunem că cele două variabile nu sunt relaționate, sunt independente.

Teoretic sună frumos, însă ce ne facem că înălțimea o măsurăm în centimetri iar coeficientul de inteligență în unități IQ. Din încurcătură ne scoate capitolul anterior. Ne amintim că există un sistem de note standardizate, numit note „z", care are proprietatea de a aduce la un numitor comun cele două variabile. Mai mult, notele „z" sunt note aditive şi multiplicative. Dacă am transforma atât înălțimea cât şi coeficientul de inteligență în note „z", am putea calcula un scor total şi am putea sa-l reprezentăm pe o distribu-

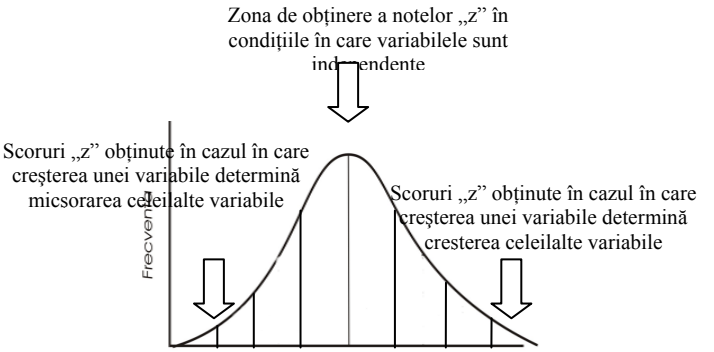

Zona de obținere a notelor „z" în condițiile în care variabilele sunt independente
Scoruri „z" obținute în cazul în care creşterea unei variabile determină micsorarea celeilalte variabile
Scoruri „z" obținute în cazul în care creşterea unei variabile determină cresterea celeilalte variabile
Frecvențe

Figura 7.9 – Reprezentarea relației dintre două variabile

ție normală. Acest lucru îl putem observa cu uşurință în figura 7.9, prezentată mai sus.

Ştim de asemenea de la semnificația statistică, faptul că testele de semnificație explică probabilitatea de obținere a unor diferențe sau relații dintre două variabile ca rezultat al unei erori de eşantionare, indicând probabilitatea ca această relație sau diferență să nu existe de fapt la nivelul populației (Dancey, şi alții, 2002). Un prag de semnificație de 0,05 înseamnă că pentru a fi respinsă ipoteza nulă, trebuie să existe mai puțin de 5% şanse ca scorurile de mai sus să fie obținute printr-o eroare de eşantionare. Să considerăm acum o ipoteză formulată în următorii termeni:

Există o legătură semnificativă între înălțimea subiecților şi coeficientul lor de inteligență.

Am formulat o ipoteză bilaterală, fără a preciza sensul acestei legături. Ipoteza nulă (*Nu există nici o legătură semnificativă între înălțimea subiecților şi coeficientul lor de inteligență*) poate fi respinsă la un prag de semnificație de 0,05 atât pentru situația în care creşterea înălțimii subiecților determină creşterea coeficientul lor de inteligență, dar şi în situația în care creşterea înălțimii duce la scăderea coeficientului de inteligență. Acest prag de 5% va fi aşadar egal distribuit la capătul din stânga şi la capătul din dreapta al distribuției. Observăm că aceste praguri corespund la aproximativ două abateri standard la stânga şi la dreapta mediei. Dacă scorurile „z" ale subiecților investigați vor fi situate în aceste zone, putem stabili legătura între înălțimea şi coeficientul de inteligență ca fiind o legătură semnificativă, la un nivel mai mic de 5% ca datele să fie obținute printr-o eroare de eşantionare.

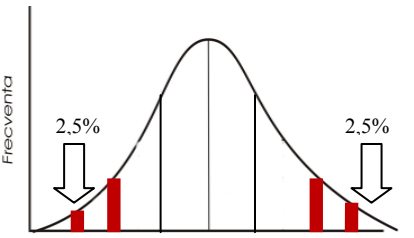


Figura 7.10 – Reprezentarea zonelor de respingere a ipotezei nule pentru ipotezele bidirecționale

Să considerăm acum cel de-al doilea caz, în care formulăm ipoteza astfel:

Există o legătură semnificativă pozitivă între înălţimea subiecţilor şi coeficientul lor de inteligenţă.

Iată o ipoteză unidirecţională care ne spune foarte clar că la creşterea înălţimii apare creşterea coeficientului de inteligenţă şi, similar, scăderea înălţimii determină scăderea coeficientului de inteligenţă. Ipoteza nulă poate fi respinsă în acest caz numai la un prag de semnificaţie mai mic de 0,05 situat în zona dreaptă a distribuţiei, la aproximativ 1,5 abateri standard la dreapta mediei, aşa cum puteţi observa în figura alăturată.

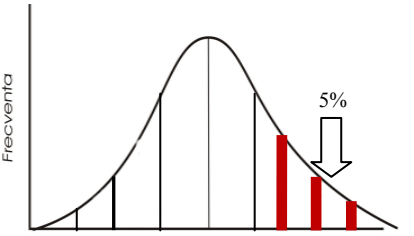


Figura 7.11 – Reprezentarea zonelor de respingere a ipotezei nule pentru ipotezele unidirecţionale

Nu suntem interesaţi ce se întâmplă dacă creşterea înălţimii implică scăderea coeficientului de inteligenţă. Putem respinge ipoteza nulă numai în cazul unei legături pozitive între cele două variabile. În mod analog se poate arăta zona de respingere a ipotezei nule pentru legăturile negative (creşterea unei variabile implică micşorarea celeilalte), aceasta fiind situată aproximativ la 1,5 abateri standard în stânga mediei.

Observăm că probabilitatea de a respinge ipoteza nulă este mult mai mare în cazul ipotezelor unidirecţionale în comparaţie cu ipotezele bidirecţionale. Există şanse mult mai mari să găsim scoruri în zona marcată din figura 7.11 în comparaţie cu zona marcată în figura 7.10. În general, dacă aţi obţinut pragul de semnificaţie în cazul unei ipoteze bidirecţionale şi doriţi să aflaţi pragul de semnificaţie pentru ipoteză unidirecţională, veţi înjumătăţi valoarea acesteia. De exemplu, dacă obţineţi valoarea 0,04 în cazul unei ipoteze bidirecţionale, corespondentul său pentru valorile unidirecţionale este de 0,02.

Prin dublarea valorii pragului de semnificație se poate obține valoarea acestuia dacă ați lucrat cu ipoteze unidirecționale şi doriți să aflați pragul de semnificație pentru o ipoteză bidirecțională.

VII.3.3 Proiectarea cercetării

Metodele statistice reprezintă un instrument de cercetare. Etapa de proiectare a cercetării nu face obiectul prezentei lucrări, existând pe piață suficientă bibliografie care tratează această temă. Abordăm totuşi, pe scurt, subiectul, din considerentele relației care există între proiectarea unei cercetări şi analiza ulterioară a datelor. Nu putem obține niciodată mai mult decât am proiectat inițial. Erorile de proiectare a cercetării induc invariabil erori de analiză şi interpretare a datelor.

Proiectarea unei cercetări necesită răspunsul la o serie de întrebări, dintre care cele mai importante sunt (Dowdy, şi alții, 2004):

- Ce tratamente sau condiții vor fi aplicate subiecților în vederea testării ipotezelor;
- Care sunt variabilele de interes şi cum vor fi acestea măsurate;
- La ce nivel de precizie vor fi efectuate măsurătorile;
- Care va fi lotul de subiecți pe care se va desfăşura cercetarea.

Orice cercetare necesită un plan de cercetare numit şi design de cercetare. În urma acestui plan se stabilesc variabilele ce vor fi analizate şi numărul de subiecți necesar. Problematica planurilor de cercetare va fi tratată în detaliu în volumul al doilea al lucrării, în timp ce proiectarea eşantionului sau a lotului de subiecți va face subiectul următorului capitol.

VII.3.4 Efectuarea observațiilor

Cel puțin la nivel teoretic, un studiu nu trebuie să conțină alte erori exceptând erorile aleatorii de măsurătoare determinate de erorile instrumentelor de măsură.

În literatura de specialitate se vehiculează frecvent termenul de **bias** vorbindu-se, chiar şi la noi în țară, de studii biasate. Biasul reprezintă tendința de afectare a măsurării unei variabile de către anumiți factori externi (Dowdy, şi alţii, 2004). Există surse de asemenea erori determinate de instrumente (erori de calibrare a instrumentului) de influența cercetătorului (proiectarea unui interviu în care există tendința sugerării răspunsurilor) de efectele de halo (tendința unei evaluator de a fi influențat de evaluările anterioare) sau chiar de condițiile atmosferice ori de subiecții înşişi.

Imediat după efectuarea observațiilor se impune inspecția datelor în vederea depistării unor valori extreme, inconsistente cu celelalte date. Inconsistențele de acest tip vor fi analizate în vederea corectării datelor sau a eliminării acestora în cazul în care metodele de corecție şi ajustare a datelor nu determină rezultate satisfăcătoare. Detalii despre aceste procedee întâlniți în capitolele anterioare.

VII.3.5 Interpretarea datelor

Procesul de interpretare a datelor debutează, după cum am afirmat, cu analiza inventarului statistic de bază. În urma acestui studiu se decide normalitatea sau lipsa de normalitate a unei distribuții, după care se aleg testele statistice necesare analizei de date.

Aplicarea testelor statistice se face în vederea verificării ipotezei nule, proces prin care decidem dacă datele experimentale sunt obținute în condițiile în care ipoteza de nul este adevărată. Dacă diferența dintre ipoteza nulă şi datele experimentale nu se datorează întâmplării (dacă pragul de semnificație

este mai mic de valoarea stabilită 0,05) putem respinge ipoteza nulă, în caz contrar ipoteza nulă nu poate fi rejectată.

Trebuie să reţineţi faptul că procedeele statistice nu demonstrează nimic. Metodele statistice descriu probabilitatea cu care un eveniment apare într-o populaţie. Poate exista doar un singur caz din zece miliarde de cazuri care să demonstreze acceptarea ipotezei nule. Prin metode statistice stabilim doar la ce nivel de încredere putem respinge ipoteza nulă şi nu putem demonstra niciodată falsitatea acesteia.

După Dancey şi Reidy (Dancey, şi alţii, 2002), majoritatea testelor statistice le utilizăm pentru a face o inferenţă de la studiul pe un eşantion spre caracteristicile populaţiei. Deoarece caracteristicile populaţiei poartă numele de parametri, testele utilizate poartă denumirea de **teste parametrice**. Pentru a putea însă utiliza testele parametrice, este necesară îndeplinirea simultană a unor condiţii axiomatice. După aceeaşi autori, (Dancey, şi alţii, 2002) există un număr de trei axiome necesare utilizării testelor parametrice:

- **Axioma normalităţii distribuţiei**. Distribuţia scorurilor obţinute de către subiecţi la variabilele analizate trebuie să fie o distribuţie normală. În absenţa unei distribuţii normale nu pot fi utilizate teste parametrice. Reamintim că utilizând testele parametrice facem asumpţia că media şi abaterea standard reprezintă indicatori reprezentativi ai tendinţei centrale, respectiv împrăştierii. Dacă această asumpţie este falsă, întreaga logică a aplicării testelor statistice va fi şi ea falsă. Nu intrăm acum în detalii, deoarece au fost prezentate anterior, ci ne rezumăm să amintim că dacă distribuţia nu este normală vor fi aplicate tehnici de normalizare a acesteia. Dacă nici în acest caz nu reuşim obţinerea unei distribuţii normale vor fi utilizate teste nonparametrice.

- **Axioma omogenităţii varianţelor.** Reprezintă o a doua asumpţie importantă care postulează că varianţa populaţiilor trebuie să fie aproximativ egală. Ca şi în cazul primei axiome, nu putem şti dacă varianţa populaţiei este aproximativ egală şi ne vom rezuma la studiul varianţei eşantionului sau lotului de cercetare.
- **Axioma scorurilor extreme.** Cea de-a treia axiomă a fost de asemenea tratată anterior şi postulează inexistenţa scorurilor extreme în distribuţia statistică. Acest lucru este evident din raţiuni ce ţin de utilizarea mediei. Nu vom mai insista asupra acestui lucru deoarece considerăm că l-am discutat suficient pe parcursul acestei lucrări.

Autorii suscitaţi nu menţionează însă o cerinţă importantă şi anume cea a **nivelului de măsură**. Evident, nu putem utiliza media şi abaterea standard decât la variabile aflate la un nivel de măsură cel puţin de interval. Este absurd să vorbim de distribuţie normală în cadrul variabilelor nominale sau ordinale.

VII.3.6 Formularea concluziilor

Formularea concluziilor reprezintă etapa finală a oricărei cercetări ştiinţifice şi rezumă întregul demers ştiinţific prezentat mai sus. Decizia luată va fi o decizie **probabilistică** şi consistentă cu datele experimentale. Dacă pragul de semnificaţie este mai mare decât pragul acceptat, atunci probabil ipoteza nulă este adevărată şi va trebui acceptată. În caz contrar putem respinge ipoteza nulă. Ipoteza experimentală nu poate fi respinsă sau acceptată. Singura decizie pe care o putem lua se referă la ipoteza de nul.

Concluzii:

- *Există patru reguli de probabilitate: probabilitatea simplă pentru evenimente egale, evenimente mutual exclusive, evenimente independente, probabilitatea condiționată;*
- *Extinderea concluziilor de la studiul unui eşantion la nivelul populației implică existența erorilor sistematice ce eşantionare;*
- *Indicatorii de sondaj aproximează parametrii populației în zona unui interval de încredere;*
- *Abaterea standard a mediilor eşantioanelor poartă numele de eroare standard de eşantionare;*
- *Metoda ştiințifică în ştiințele socio-umane presupune parcurgerea unui număr de şase etape:*
    - *Enunțarea problemei;*
    - *Formularea ipotezelor;*
    - *Proiectarea cercetării;*
    - *Efectuarea observațiilor;*
    - *Interpretarea datelor;*
    - *Formularea concluziilor;*
- *Enunțarea problemei presupune formularea obiectivelor generale şi specifice, în general în urma unui demers metaanalitic;*
- *Specificarea ipotezelor face referire la ipoteza alternativă şi la ipoteza nulă;*
- *Ipotezele pot fi unidirecționale sau bidirecționale;*
- *Acceptarea sau respingerea ipotezei nule se face prin referire la pragul de semnificație statistică;*
- *Aplicarea testelor parametrice se face în baza unor axiome;*
- *Orice concluzie şi decizie luată în urma aplicării metodei statistice este o decizie probabilistică şi nu certă.*

Întrebări şi exerciții:

- *Care dintre următoarele evenimente au o probabilitate de apariție 0 şi care 1 sau apropiată de 1?*
    - *După zi urmează noapte;*
    - *Preşedintele României spune întotdeauna adevărul;*
    - *Veți găsi un milion de euro în paginile acestei cărți;*
    - *Când vă este sete, beți apă;*
    - *Stăpâniți bazele statisticii citind această carte;*
- *Exprimați următoarele probabilități sub formă procentuală:*
    - *0,25;*
    - *0,13;*
    - *0,33;*

- *1/9;*
- *6/8;*

- *Exprimați următoarele probabilități sub formă zecimală:*
  - *1/7;*
  - *3/15;*
  - *9/245;*
  - *84%;*
  - *39%;*

VIII. EŞANTIONARE ŞI REPREZENTATIVITATE

În acest capitol se va discuta despre:

- Conceptele populaţie şi eşantion;
- Tipuri şi surse de date;
- Caracteristicile unui eşantion;
- Tipuri de eşantioane;
- Tehnici de construcţie a unui eşantion;

După parcurgerea capitolului, cursanţii vor fi capabili să:

- Înţeleagă conceptul de populaţie;
- Identifice sursele de colectare a datelor;
- Înţeleagă principiul statistic al extragerii eşantionului dintr-o populaţie;
- Facă diferenţa între indicatori şi parametri;
- Aleagă tipul de eşantion adecvat cercetării;
- Stabilească şi să caracterizeze proprietăţile unui eşantion
- Realizeze eşantioane utilizând SPSS for Windows

Pe parcursul acestui volum am abordat de multe ori problematica eşantioanelor şi am discutat sumar despre o serie de metode de eşantionare. A venit momentul să tratăm în detaliu aceste elemente, deoarece foarte puţine lucrări de statistică socială le abordează la un nivel comun, iar importanţa stabilirii unui eşantion reprezentativ condiţionează în mod direct valoarea cercetării. Dacă la nivelul unor studii specifice eşantionul nu primează, atunci când vorbim de elaborarea unor teste psihologice ori despre desfăşurarea cercetărilor a căror rezultate vor caracteriza populaţia, procesul de eşantionare dobândeşte o valoare specială.

Vom relua, pentru început, o serie de lucruri deja cunoscute, referitoare la populaţie şi eşantion, ulterior urmând să aprofundăm atât tipurile de eşantioane cât şi metodele specifice de construcţie ale acestora.

VIII.1 Populația

Definiția din capitolul al treilea afirma că populația reprezintă totalitatea obiectelor, de un anumit tip, existente într-un spațiu sau teritoriu, la un moment dat. O definiție mai largă susține că **o populație reprezintă o colecție naturală, geografică sau politică de persoane, animale, plante sau obiecte** (Dowdy, şi alţii, 2004). Definiția subliniază că o populație reprezintă, de fapt, o colecție fizică. Fiind o colecție fizică, orice populație se caracterizează prin anumite proprietăți pe care le denumim generic **parametri**. Termenul de populație este, de fapt, o abstracțiune. Nu putem concepe populația fără să facem apel la aceste proprietăți prin care se defineşte. Dacă ne referim la populația bărbaților, implicit am definit un parametru al acesteia şi anume sexul. Dacă vorbim de populația înălțimii bărbaților, avem în vedere doi parametri şi anume înălțimea şi sexul. Similar, înălțimea bărbaților din România presupune trei parametri: locația, sexul şi înălțimea. Toți aceşti parametri se numesc **variabile de interes ale cercetării** şi permit înregistrarea unor valori la nivelul acestora.

Într-o cercetare, odată stabilite variabilele (parametrii) de interes, trebuie să acordăm multă atenție în momentul în care extindem concluziile la nivelul populației. Vom reține faptul că o populație se caracterizează prin aceste variabilele de interes. Nu putem face o afirmație de genul „înălțimea bărbaților este de 178 de centimetri", în condițiile în care am efectuat un studiu pe populația din România. Nu am desfăşurat o cercetare prin care să investigăm înălțimea bărbaților din lume. O formulă adecvată ar fi „înălțimea bărbaților din România este de 178 de centimetri".

Populația odată definită, apare necesitatea înregistrării valorilor pentru variabilele de interes. O metodă de investigație exhaustivă – care-şi propune investigarea tuturor elementelor dintr-o populație – este recensământul.

Recensământul este o metodă de observare totală, cu caracter periodic, care surprinde un fenomen în mod static. Este una din cele mai vechi

metode de observare şi asigură o fotografiere, o surprindere a unui fenomen într-un anumit moment de timp. În mod particular, recensământul populației este un proces de culegere, prelucrare şi publicare a datelor demografice, economice şi sociale, la un timp specificat şi valabile pentru toate persoanele din țara respectivă sau de pe un teritoriu delimitat. O formă particulară a recensământului o reprezintă **referendumul,** ca demers similar de investigare a opiniilor populației.

O asemenea metodă de colectare a informațiilor este extrem de costisitoare. De multe ori nu este practic (uneori chiar imposibil) să observăm toate valorile pe care le poate lua o variabilă la nivelul populației.

VIII.2 Surse de date

Odată stabilite variabilele de interes şi definită populația, se pune problema colectării datelor necesare. Care sunt însă sursele de date? După Curwin şi Slater (Curwin, şi alţii, 1996), datele pot proveni din surse principale şi/sau din surse secundare.

VIII.2.1 Sursele principale de date

Reprezintă surse care permit colectarea directă de observații noi. Metodele prin care putem efectua un asemenea demers includ observația, discuțiile individuale sau de grup, experimentele, chestionarele, testele, utilizarea unor aparate specifice etc. Colectarea datelor la acest nivel poate fi costisitoare atât din punct de vedere financiar cât şi sub aspectul duratei, însă avem certitudinea unor date reale şi valide. La nivelul elementelor de populație, cercetarea poate presupune o explorare exhaustivă, ca în cazul recensământului, sau poate implica utilizarea eşantioanelor. În mod curent, cercetările folosesc, de cele mai multe ori, surse principale de date.

VIII.2.2 Sursele secundare de date

Reprezintă date provenite din cercetări deja efectuate. Costurile necesare procurării acestor date sunt semnificativ mai reduse, uneori informaţiile sunt de o înaltă valoare ştiinţifică, însă dezavantajul principal este acela că nu putem avea controlul asupra veridicităţii şi validităţii acestora. Astfel, datele pot proveni din anuarele statistice ale unor ţări sau teritorii, din rapoartele unor organizaţii sau instituţii, din cercetări existente sau din baze de date publice ori private. Metaanaliza, ca metodă ştiinţifică, utilizează predominant asemenea surse de date.

VIII.3 Eşantionul

Exceptând recensământul, toate celelalte metode implică utilizarea unei mici părţi dintr-o populaţie în vederea colectării datelor. O asemenea colectivitate statistică poartă numele de eşantion şi reprezintă **un subset de elemente din populaţie, care însă păstrează caracteristicile populaţiei din care provine**. O cercetare efectuată pe un eşantion, induce, după cum am arătat, o serie de erori care afectează rezultatele finale. Aceste erori apar ca urmare a faptului că noi nu culegem informaţii de la întreaga populaţie statistică, ci doar de la elementele cuprinse în eşantion. O cercetare efectuată pe un eşantion poartă numele de **cercetare selectivă**, spre deosebire de recensământ care se referă la **cercetări exhaustive**.

Vă mai amintiţi exemplul cu oala de fasole? Atunci am extras o lingură de boabe de fasole şi am decis că fasolele sunt fierte. Exact acesta este principul eşantionării. Dacă am fi luat o singură boabă de fasole, am fi putut face această afirmaţie? Eu cred că nu, deoarece ar fi existat prea multe şanse să aruncăm apoi oala. Dar dacă am fi mâncat întreaga oală? Atunci am fi ştiut sigur, însă nu am mai fi avut cu ce să facem mâncarea. În plus, gândiţi-vă ce costuri aţi fi suportat cu sistemul digestiv… Dacă în oală aţi fi amestecat fasole de 10 ani şi fasole de 1 an? Probabil că decizia ar fi fost corectă, însă aţi

avea multe şanse să vă rupeți uneori dinții. Dacă ați fi gustat din oala de mazăre? Desigur că nu, pentru că v-ați fi referit la cu totul altă populație.

Iată că un eşantion trebuie să respecte caracteristicile populației din care face parte. Aceasta este calitatea esențială a unui eşantion şi poartă numele de **reprezentativitate.** Reprezentativitatea este, aşadar, **capacitatea eşantionului de a reproduce cât mai fidel structurile şi caracteristicile populației din care a fost extras** (Rotariu, şi alții, 2006). Dacă valorile proprietăților la nivelul populației poartă numele de parametri, la nivelul eşantioanelor ele se numesc **indicatori** sau **statistici**. În mod normal noi nu cunoaştem valorile parametrilor, valoarea exactă pe care o are caracteristica respectivă la nivelul populației, ci încercăm s-o estimăm prin intermediul valorii obținute la nivelul eşantionului. Legat de acest aspect, trebuie însă să rețineți două lucruri importante:

- Între valoarea unui indicator şi valoarea unui parametru există întotdeauna o diferență, numită **eroare** sau **eroare statistică**. Sensul termenului de eroare nu este acela comun, de greşeală. Dacă am măsura toți bărbații din România şi am obține 179 centimetri pentru parametrul înălțime medie, în timp ce un alt studiu, efectuat pe 200 de bărbați obține înălțimea medie de 178,7 centimetri, diferența de 0,3 centimetri reprezintă tocmai eroarea statistică;

- Pentru ca un studiu să fie valid, trebuie ca eroarea (diferența dintre valoarea parametrului şi cea a indicatorului) să nu depăşească un anumit prag. Acest prag poartă numele de **nivel de încredere** şi reprezintă şansa ca eroarea real comisă să nu depăşească limita erorii statistice. Cu alte cuvinte, şansa ca valorile obținute să nu se datoreze unei erori de eşantionare.

Aceste lucruri le-am abordat pe larg în capitolul anterior şi nu le vom mai discuta aici deoarece consider că au fost bine înțelese. Totuşi, vom prelua

o serie de principii ale reprezentativității eşantionului, aşa cum le prezintă T. Rotariu în lucrarea sa (Rotariu, şi alţii, 2006):

- Reprezentativitatea unui eşantion trebuie întotdeauna evaluată prin eroarea statistică şi nivelul de încredere, deoarece în absența specificării nivelului de încredere nu putem fi siguri că eroarea comisă este mai mică decât eroarea statistică;
- Reprezentativitatea unui eşantion, ca şi concept statistic, la fel ca şi populația, este un termen abstract. Nu putem vorbi de reprezentativitate decât în raport cu o caracteristică dată (vârstă, gen biologic, înălțime, coeficient de inteligență etc.);
- Există diferite grade de reprezentativitate a unui eşantion. Un eşantion poate fi mai reprezentativ decât altul, în condițiile în care la un acelaşi nivel de încredere eroarea este mai mică, sau la aceeaşi eroare, nivelul de încredere este mai ridicat;

Din cele expuse mai sus putem concluziona că reprezentativitatea unui eşantion depinde de **caracteristicile populației** ce va fi studiată, de **mărimea eşantionului** şi de **procedura folosită** (Rotariu, şi alţii, 2006).

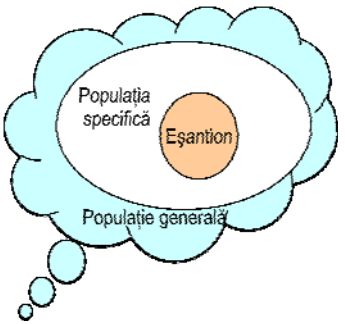


Figura 8.1 – Relația dintre populația generală, variabilele de interes şi eşantion

Principul de bază al eşantionării este expus succint în figura alăturată. Din populația generală se construieşte populația țină în funcție de variabilele de interes necesare cercetării.

În funcție de nivelul de precizie ales, se determină mărimea eşantionului, în baza populației țintă, astfel încât să se poată stabili eroarea maximă şi nivelul de încredere.

Desigur, problema se pune la nivelul stabilirii mărimii eşantionului, astfel încât acesta să fie reprezentativ.

VIII.3.1 Mărimea eşantionului

V-aţi pus în mod cert întrebarea, în nenumărate rânduri, cum poate un sondaj să vorbească în numele oamenilor. Sau cum pot cei care organizează sondaje să ghicească cine va ieşi preşedinte, evident dacă sondajele sunt corecte. Dacă nimeni nu m-a întrebat nimic, cum poate cineva să facă o afirmaţie valabilă şi pentru mine? Este o întrebare pertinentă, însă vă daţi seama ce s-ar întâmpla şi ce ar implica investigarea întregii populaţii la fiecare cercetare? Nu numai că ar fi prohibitive costurile, însă nimeni nu ar putea da rezultatele în timp util şi, mai ales, nu ar fi necesar.

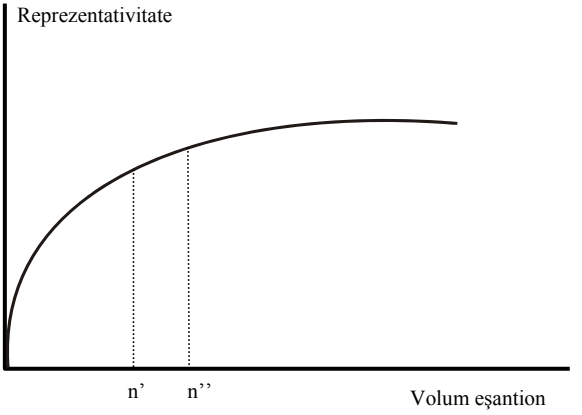


Figura 8.2 – Relaţia dintre volumul eşantionului şi reprezentativitate

În mod cert, reprezentativitatea unui eşantion creşte odată cu creşterea numărului de elemente şi atinge maximum la un volum egal cu cel al populaţiei. Totuşi, relaţia dintre reprezentativitate şi volumul eşantionului nu este una liniară, ci una curbiliniară. În figura alăturată, putem observa că în cazul unui eşantion format din zero persoane, reprezentativitatea lui este şi ea nulă. Cu cât creşte numărul de persoane, cu atât creşte şi reprezentativitatea, însă până la un punct situat între *n'* şi *n''*. După acest nivel, reprezentativitatea eşantionului creşte foarte puţin pe măsură ce creşte volumul acestuia, astfel încât diferenţele vor fi nesemnificative. Toată problema este să determinăm punctul pentru care reprezentativitatea este maximă, punct situat între *n'* şi *n''* care va determina mărimea (*n*) a eşantionului. Un volum al eşantionului mai mic de *n'* duce la un eşantion nereprezentativ, iar un volum al eşantionului mai mare de *n''* determină cheltuieli nejustificate.

Un fapt importat pe care dorim să-l abordăm, este acela că, **reprezentativitatea eşantionului nu depinde de mărimea populației**. Fie că desfăşurăm un studiu în România, în India sau în Statele Unite, un eşantion de 500 de persoane va avea aceeaşi reprezentativitate.

VIII.3.2 Caracteristicile populației

Ce ne facem în situația în care avem în oală fasole de 1 an şi fasole de 10 ani? Iată că în termeni statistici, populația de fasole din oala dumneavoastră este formată, de fapt, din două subpopulații: cea de 10 ani şi cea de 1 an. Deşi lingura de fasole pe care o gust este reprezentativă la nivelul populației, ea nu mai este reprezentativă la nivelul subpopulațiilor. Sunt de fapt două straturi în populația noastră, fiecare dintre aceste straturi având ponderi diferite. Reprezentativitatea nu are de-a face cu această proporție a straturilor în raport cu volumul populației, astfel încât, în această situație, eşantionul boabelor de fasole din lingura mea nu mai este reprezentativ la nivelul celor două subpopulații de boabe de fasole. Ce este de făcut în acest caz? Dacă cunoaştem ponderea straturilor, putem folosi o procedură specifică de eşantionare. Dacă nu, în general gustăm mai multe linguri, adică mărim efectivul eşantionului.

VIII.3.3 Procedee de eşantionare

Metodele şi tehnicile de eşantionare au fost abordate şi în capitolul al treilea al acestui volum. În cadrul acestei secțiuni, vom studia în detaliu fiecare tehnică, specificând, atunci când este cazul, modalitatea de calcul. Cum aflăm totuşi cea mai bună metodă de eşantionare? În figura de mai jos aveți un algoritm de decizie în vederea alegerii procedurii adecvate de eşantionare.

În practică, se face distincția între procedeele de eşantionare aleatoare şi procedeele de eşantionare nealeatoare. Din punct de vedere statistic însă, orice procedeu de eşantionare nealeatoriu nu poate fi tratat prin prisma teoriei probabilităților, fiind, prin urmare, eliminat din categoria metodelor de eşan-

tionare. Când discutăm de reprezentativitate, vom aborda strict procedeele de eşantionare aleatorii. **O procedură de eşantionare este aleatoare în momentul în care fiecare individ din populaţie are o şansă calculabilă de a fi ales în eşantion**.

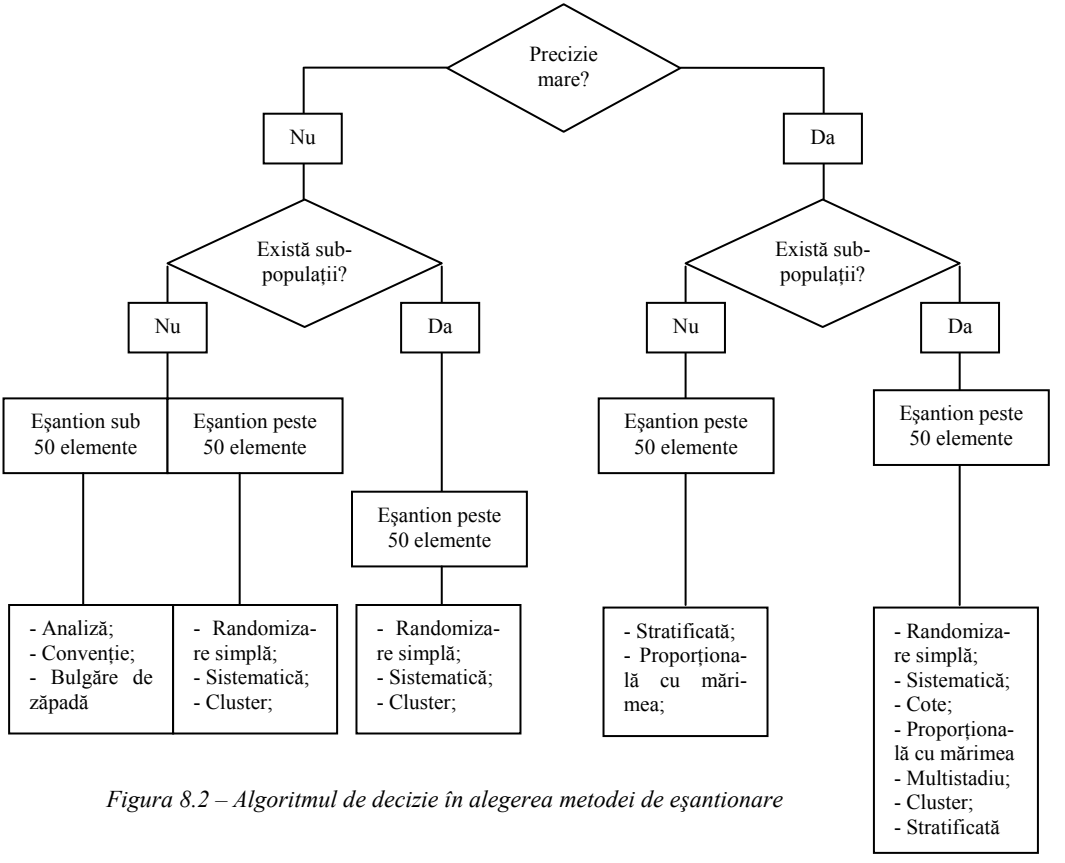


Figura 8.2 – Algoritmul de decizie în alegerea metodei de eşantionare

VIII.3.3.1 Stabilirea eşantionului prin convenţie

Este un procedeu nealeatoriu de stabilire a unui eşantion, bazat pe voluntariat sau pe cazuri specifice, analizate de cercetător. De exemplu, situaţia în care un psihoterapeut îşi construieşte un studiu ştiinţific bazat pe clienţii

săi, pe care îi tratează, ori în cazul cercetărilor efectuate pe studenți, în cadrul unei universități. Nu putem vorbi aici de reprezentativitate, iar rezultatele obținute nu pot fi extrapolate la nivelul întregii populații. Un asemenea studiu, datorită voluntariatului, este puternic biasat, iar rezultatele vor fi privite cu mare scepticism. Avantajul îl reprezintă posibilitatea de colectare facilă şi rapidă a datelor, iar cu cât grupul este mai mare, cu atât rezultatele vor fi mai importante. Într-o cercetare în care se folosesc asemenea subiecți, în general, elementele se raportează ca lot de cercetare sau lot de studiu şi nu ca eşantion, tocmai pentru a specifica lipsa oricăror tehnici de eşantionare.

VIII.3.4 Stabilirea eşantionului prin analiză

Reprezintă un alt procedeu nealeatoriu de colectare a datelor, bazat pe alegerea deliberată a cazurilor. Nici în această situație nu putem vorbi de eşantion, preferându-se termenul de lot de cercetare. Erorile statistice sunt şi în acest caz foarte mari şi nu se pot extinde cunoştințele la nivelul întregii populații, iar cercetarea prezintă probleme de credibilitate. Se foloseşte în general pentru studiile de caz sau în condițiile unor populații cu un comportament bine cunoscut.

Cele două metode prezentate mai sus presupun utilizarea unor loturi de cercetare de mici dimensiuni (în general sub 50 de cazuri) şi le întâlnim frecvent în lucrări specifice domeniului psihologic. Din nefericire, concluziile rezultate în urma unor asemenea studii au o valoare limitată la nivelul lotului cercetat, orice posibilitate de generalizare fiind statistic exclusă.

VIII.3.4.1 Metoda bulgărelui de zăpadă

Este o tehnică folosită destul de frecvent în cercetările socio-umane, în special în sociologie şi se bazează pe metode multi-level. O persoană care răspunde, recomandă în acelaşi timp o altă persoană care va răspunde (un prieten, o cunoştință etc.). Aceasta, la rândul ei, recomandă o altă persoană şi aşa mai departe. Cercetările de acest tip fixează, de obicei, o dată limită până

la care se pot primi răspunsurile subiecților. În mod cert, metoda de alegere a subiecților nu este una aleatorie şi prezintă limitările prezentate mai sus. Avantajul este acela al costurilor reduse şi al creşterii exponențiale a volumului de date. Într-un timp scurt se pot colecta foarte multe date, însă cercetările nu prezintă relevanță, datorită caracterului nealeatoriu şi a faptului că nu există un control al răspunsurilor. Fiecare respondent poate să-şi exprime punctul său real de vedere sau un punct de vedere fals.

VIII.3.4.2 Eşantionarea simplă aleatoare

Reprezintă clasicul procedeu al loteriei sau al tragerii la sorți, fiind considerată cea mai simplă metodă de eşantionare. Nu presupune alte operațiuni prealabile, iar prin această metodă fiecare individ dintr-o populație are şansa de a fi ales în eşantion. Când eram copil, obişnuiam să joc fotbal în echipa de cartier. Bineînțeles că problema era cine să fie portar, deoarece nici unul dintre noi nu dorea. Deoarece ne strângeam cam 30-40 de copii, singura soluție era să tragem la sorți, astfel încât făceam un număr de bilețele egal cu numărul nostru, iar pe un singur bilețel scriam „portar". Amestecam apoi bilețelele într-o căciulă şi fiecare trăgea câte unul. Ghinionistul stătea în poartă şi, evident, au fost suficiente ocazii în care cu sufletul trist aveam această „onoare". Nu ştiam atunci că aplic un procedeu statistic de eşantionare, numit eşantionarea simplă aleatoare.

Nu cred că este cazul să mai menționez faptul că această metodă este o modalitate aleatorie de eşantionare, prin care se poate estima un parametru al populației, se poate calcula eroarea de eşantionare, are un design simplu şi eficient, iar interpretarea este foarte uşoară.

Strict matematic, pentru a construi un eşantion aleatoriu se folosesc tabele de numere aleatoare, formate din şiruri lungi de numere alese la întâmplare, atât sub aspectul numerelor în sine cât şi sub cel al poziției în care se află în tabel. Metodele informatizate pot folosi, în acelaşi scop, generatoa-

re de numere aleatorii. Stabilirea unui eşantion pe baza eşantionării aleatorii simple presupune următoarele etape:

- **Stabilirea populației țintă**. Considerăm că dorim să studiem înălțimea bărbaților din România. Populația va fi formată din totalitatea bărbaților din România, persoane de sex masculin cu vârsta de peste 18 ani. Să presupunem că numărul acestora este de 8 milioane (*N=8000000*);
- **Stabilirea bazei de eşantionare**. Prin „bază de eşantionare" înțelegem o listă exhaustivă a populației țintă. Avem nevoie, prin urmare, de o listă a tuturor bărbaților din România pe care o numerotăm succesiv de la 1 la 8 milioane. Practic se obține inițial o listă ordonată alfabetic, se generează apoi, aleatoriu, pentru fiecare înregistrare din listă, un număr şi se reordonează lista după numărul generat. Astfel, amestecăm baza de eşantionare, eliminând criteriul alfabetic. Pe noua listă amestecată, generăm apoi numere succesive de la 1 la 8 milioane;
- **Stabilirea mărimii eşantionului**. Prin procedee de calcul pe care le vom aborda în următorul subcapitol, stabilim dimensiunea eşantionului. În situația noastră, presupunem că alegem un eşantion de 1000 de persoane (*n=1000*);
- **Generarea numerelor aleatoare**. Se generează pe computer 1000 de numere aleatoare sau se utilizează tabelul cu numere aleatoare;
- **Extragerea eşantionului**. Se citeşte primul număr aleatoriu generat sau primul număr din tabelul de numere aleatoare. Acel număr reprezintă numărul de ordine al subiectului din baza de eşantionare care va fi inclus în eşantion. Se trece apoi la al doilea număr aleatoriu, acesta reprezentând numărul de

ordine al celui de-al doilea subiect şi se continuă astfel până la completarea efectivului (1000 de subiecți) eşantionului.

Deşi extrem de simplă, metoda prezintă o serie de limitări serioase. Principala problemă o constituie obținerea bazei de eşantionare. Cum găsim noi lista cu numele, prenumele şi adresa tuturor bărbaților din România? Unii cercetători folosesc cărțile de telefon şi generează aleatoriu volumul eşantionului în baza numerelor de telefon. Se pune însă problema dacă toată populația are telefon. Cei care nu au telefon sau nu figurează în cartea de telefon sunt automat excluşi din baza de eşantionare, ceea ce duce la eliminarea caracterului aleatoriu.

Pe de altă parte, presupunând că am obținut baza de eşantionare, toate persoanele din eşantion sunt accesibile? Dacă unii sunt plecați în străinătate, au murit, s-au mutat ori nu pot fi contactați, ce se întâmplă? Se mai respectă reprezentativitatea eşantionului? Vă las pe dumneavoastră să decideți acest lucru. În general, cu cât volumul populației țintă este mai mare, cu atât această metodă este mai dificil de implementat.

VIII.3.4.3 Eşantionarea stratificată

Iată şi metoda utilizată în verificarea fasolelor de 1 an şi a celor de 10 ani. Eşantionarea stratificată este considerată superioară eşantionării aleatorii simple deoarece, în condițiile în care la nivelul populației țintă există mai multe straturi (subpopulații), se va putea reproduce ponderea fiecărui strat în eşantionul final.

Se consideră că dintre două eşantioane de volum egal, cel produs prin stratificare va avea o reprezentativitate mai bună în comparație cu cel produs prin eşantionarea aleatorie simplă (Rotariu, şi alții, 2006).

Procedura concretă de lucru constă în următoarele etape principale:

- **Stabilirea variabilelor de stratificare.** Stratificarea se poate face după o singură variabilă (stratificare simplă) sau după mai

multe variabile (stratificare multiplă). Spre exemplu, ne interesează consumul mediu zilnic de ţigări pentru populaţia din România. Care ar fi variabilele de stratificare? Femeile consumă oare mai multe ţigări în comparaţie cu bărbaţii? Grupa de vârstă influenţează consumul de ţigări? Cei din mediul rural fumează mai mult sau mai puţin în comparaţie cu cei din mediul urban? Nu există o regulă de stabilire a variabilelor de stratificare, acestea fiind alese în funcţie de scopul cercetării şi de rolul pe care îl au în legătură cu tema cercetată. Întotdeauna însă, suma cazurilor din fiecare strat trebuie să redea totalul populaţiei. Dacă ne interesează numai straturile determinate de sex, efectuăm o **stratificare simplă**. Dacă însă luăm în considerare sexul şi mediul de proveniență, avem o **stratificare multiplă**.

Să luăm un exemplul, prin care dorim să studiem consumul mediu de ţigări din judeţul Blagosloveşti, cu o populaţie totală de 450.380 de persoane. Ne interesează o stratificare în funcţie de variabilele sex şi mediu de proveniență. Putem obţine, în această etapă, următoarea structură a populaţiei, prezentată în tabelul de mai jos:

Tabel 8.1 – Structura populaţiei pentru cele două variabile de stratificare în judeţul Blagosloveşti

Bărbaţi		**Femei**	
Rural	**Urban**	**Rural**	**Urban**
120.215	80.360	113.456	136.394
(26,69%)	(17,84%)	(25,19%)	(30,28%)
200575 (44,53%)		**249805 (55,47%)**	
450.380 (100%)			

Problema pe care ne-o punem se referă la proporția (procentul) din populația țintă reprezentată de fiecare strat. Pentru a afla acest lucru, nu ne rămâne decât să aplicăm regula de trei simplă:

Dacă 450.380 înseamnă 100%

Atunci 120.215 înseamnă X%

$$X = \frac{120.215 \times 100}{450.380} = \frac{1.202.150}{450.380} = 26{,}69\%$$

În mod similar vom efectua calculele pentru toate cele patru secțiuni. Remarcăm că totalul subiecților din fiecare variabilă de stratificare reprezintă întreaga populație țintă.

- **Stabilirea bazei de eşantionare.** Se realizează la fel ca în cazul eşantionării aleatorii simple. Avem nevoie de lista completă a tuturor persoanelor din județul Blagosloveşti, prelucrată după modalitatea prezentată în subcapitolul anterior. Singura diferență este că baza de eşantionare va fi organizată pe secțiuni diferite, în funcție de variabilele de stratificare. În cazul nostru, vom avea patru secțiuni ale bazei de eşantionare şi anume: bărbați din mediul urban, femei din mediul urban, bărbați din mediul rural şi femei din mediul rural. (*N=450.380*);

- **Stabilirea mărimii eşantionului.** Se face în urma unor calcule specifice, adecvate acestui tip de eşantionare, calcule pe care le vom prezenta într-un capitol ulterior. În cazul nostru am ales un eşantion de 300 de persoane (*n=300*);

- **Stabilirea compoziției eşantionului.** Reprezintă etapa prin intermediul căreia ne asigurăm că proporția din populație o regăsim şi în eşantion. Vom folosi de asemenea regula de trei simplă.

Dacă 300 înseamnă 100%

Atunci X înseamnă 26,69%

$$X = \frac{26{,}69 \times 300}{100} = \frac{8007}{100} = 80\ cazuri$$

Tabel 8.2 – Structura eşantionului pe cele două variabile de stratificare în judeţul Blagoslovești

Bărbați		**Femei**	
Rural	**Urban**	**Rural**	**Urban**
80 (26,69%)	54 (17,84%)	76 (25,19%)	91 (30,28%)
134 (44,53%)		**167 (55,47%)**	
301 (100%)			

Observăm că structura eşantionului selectat reproduce fidel structura populației din care a fost extras, pentru fiecare dintre cele două variabile de stratificare. Desigur, în loc de 300 de subiecți au rezultat 301 subiecți, datorită rotunjirilor apărute prin calcul. Vom păstra în cele din urmă această valoare obținută.

- **Generarea numerelor aleatoare.** Se realizează similar eşantionării aleatorii simple pentru fiecare bază de eşantionare. Se vor genera 80 de numere aleatoare pentru bărbații din mediul rural, 54 de numere aleatoare pentru bărbații din mediul urban, 76 de numere aleatoare pentru femeile din mediul rural şi 91 de numere aleatoare pentru femeile din mediul urban;
- **Extragerea eşantionului.** Numerele aleatoare generate în etapa anterioară reprezintă numerele de ordine pentru subiecții din cele patru baze de eşantionare. Eşantionul final va fi compus din 301 persoane, aparținând tuturor celor patru categorii.

Deşi mai laborioasă decât eşantionarea aleatorie simplă, eşantionarea stratificată are avantajul unei mai bune reprezentativități.

Exercițiu:

Calculați câte boabe de fasole de 10 ani şi câte boabe de fasole de 1 an trebuie să aveți în lingură, cunoscând că în oală se află 854 de boabe de fasole de 1 an şi 362 de boabe de fasole de 10 ani.

Deşi aparent caracterul aleatoriu nu se păstrează datorită creării ponderilor pe straturi, această alegere a subeşantioanelor asigură tuturor indivizilor o probabilitate egală de a intra în eşantion. O problemă poate să apară în situația în care ponderea unui strat este foarte mică în populație. Dacă considerăm variabila de eşantionare „nivelul de inteligență", există posibilitatea ca grupa geniilor sau a idioților să aibă foarte puține cazuri în populație (să spunem 10). În această situație, în eşantion ar trebui să includem 0,003 (un exemplu, cifra nu a rezultat din calcul) persoane, ceea ce este imposibil. Asemenea situații impun fie o supra-reprezentare (adică includerea mai multor cazuri în eşantion decât proporția din populație) urmată apoi, în decursul analizei indicatorilor statistici, fie de ponderarea valorilor parțiale la nivelul stratului respectiv, fie de includerea acelui strat într-un alt strat supraordonat sau subordonat (de exemplu, includerea idioților în categoria imbecililor şi redenumirea stratului „idioți-imbecili").

VIII.3.4.4 Eşantionarea multistadială

Deşi cu o reprezentativitate mai redusă în comparație cu eşantionarea aleatorie simplă, această metodă este intens utilizată în cercetările sociologice din rațiuni de eficiență practică şi cost. Se foloseşte în special atunci când volumul populației este foarte mare (studii la nivel național), costurile cercetării sunt limitate şi se urmăreşte obținerea rapidă a unor date.

Principiul de bază al acestei eşantionări este acela al proximităţii spaţiale şi al faptului că persoanele pot fi privite ca indivizi aparţinând unor grupuri. Eşantionarea multistadială presupune parcurgerea următoarelor etape:

- **Stabilirea stadiilor (grupurilor).** Este etapa prin intermediul căreia definim grupurile din care vor fi extrase eşantioanele. Dacă am dori să investigăm, la nivel naţional, consumul mediu de ţigări pe zi, am putea efectua o eşantionare aleatorie simplă sau stratificată. Ce ne facem însă cu operatorii care trebuie să administreze 100 de chestionare unor persoane din câteva judeţe, puternic dispersate spaţial. Efortul şi costurile ar fi substanţiale. Prin eşantionarea multistadială ne putem referi, spre exemplu, la organizarea teritorială. România este organizată pe judeţe, la nivelul judeţelor se află municipii, oraşe, comune şi sate, municipiile sunt organizate pe cartiere etc. Grupurile pe care le putem folosi ar fi judeţele, oraşele, comunele şi satele.

- **Stabilirea eşantionului de rangul I.** Presupune extragerea unui eşantion din grupul de prim rang. În exemplul nostru, grupul de prim rang este reprezentat de judeţe. În România există un număr de 42 de judeţe, din care, prin procedee aleatorii, vom extrage, să spunem, un eşantion de 5 judeţe (Constanţa, Bacău, Timiş, Satu Mare şi Cluj). Aceste cinci judeţe formează populaţia ţintă pentru următorul stadiu.

- **Stabilirea eşantionului de rangul II.** Presupune extragerea, din judeţele selectate în etapa anterioară, a unui număr de localităţi. În cazul nostru, judeţul Constanţa are un număr de 69 de localităţi, judeţul Bacău – 88 de localităţi, judeţul Timiş – 92 de localităţi, judeţul Satu Mare – 60 de localităţi şi judeţul Cluj – 83 de localităţi. În total, avem un număr de 392 de loca-

lități, din care calculăm un eşantion de 17 localități pe care le extragem prin procedee aleatorii. Vor rezulta 3 localități din Bacău (Căiuți, Dărmăneşti şi Sascut), 5 localități din Constanța (Ostrov, Dobromir, Saraiu, Medgidia şi Techirghiol), 2 localități din Satu Mare (Vama şi Carei), 4 localități din Cluj (Mihai Viteazu, Ciucea, Săcuieu şi Vad) şi 3 localități din Timiş (Tomeşti, Cenad şi Pişchia). Aceste localități vor reprezenta baza de eşantionare pentru următorul stadiu.

- **Stabilirea eşantionului de rangul III.** Putem continua cu grupurile până la nivel de individ sau putem include, în această etapă, un alt grup care va funcționa ca şi stadiu de eşantionare. Mă voi rezuma acum la a considera acest stadiu ca fiind ultimul şi vom proceda la extragerea eşantionului final. Baza de eşantionare va fi constituită din totalul populației celor 17 localități, sub forma binecunoscutei liste exhaustive de populație. Presupunem că în cele 17 localități trăiesc 1.854.360 de persoane (*N=1.854.360*)

- **Stabilirea mărimii eşantionului**. Prin procedee de calcul pe care le vom aborda în următorul subcapitol, stabilim dimensiunea eşantionului. În situația noastră, presupunem că alegem un eşantion de 200 de persoane (*n=200*);

- **Generarea numerelor aleatoare**. Se generează pe computer 200 de numere aleatoare sau se utilizează tabelul cu numere aleatoare;

- **Extragerea eşantionului**. Se citeşte primul număr aleatoriu generat sau primul număr din tabelul de numere aleatoare. Acel număr reprezintă numărul de ordine al subiectului din baza de eşantionare care va fi inclus în eşantion. Se trece apoi la al doilea număr aleatoriu, acesta reprezentând numărul de

ordine al celui de-al doilea subiect şi se continuă astfel până la completarea efectivului de 200 de subiecți solicitat de eşantion.

Deşi această metodă de eşantionare permite construirea unor eşantioane aflate în proximitate spațială, reducând costurile de cercetare şi facilitând achiziția datelor, se poate întâmpla ca la nivelurile de rang înalt, elementele să fie selectate numai dintr-o anumită zonă (spre exemplu 3 județe din Moldova şi 2 din Ardeal). În acest caz, o mare parte a teritoriului rămâne neacoperită, ceea ce determină probleme serioase de reprezentativitate. Pentru a evita acest lucru se combină eşantionarea multistadială cu eşantionarea stratificată.

O altă problemă care se poate pune este cea a inegalității grupurilor. La un anumit nivel de eşantionare, se creează, într-adevăr, o probabilitate egală de alegere a tuturor elementelor, însă între Covasna cu 45 de localități şi Mureş cu 111 localități există o importantă diferență de mărime. Pentru a evita acest lucru, în general se ponderează probabilitatea de extragere a unui grup, proporțional cu mărimea acestuia.

Un alt dezavantaj al acestei metode îl reprezintă calculele laborioase. De obicei resursele economisite prin colectarea datelor sunt pe deplin folosite la proiectarea eşantionului, în condițiile unei reprezentativități satisfăcătoare.

VIII.3.4.5 Eşantionarea pe clusteri

Reprezintă o metodă ieftină de eşantionare, însă costurile scăzute sunt compensate pe deplin de erori de eşantionare destul de mari. În plus, dacă volumul clusterului este mare, cresc şi cheltuielile de colectare a datelor. Acest tip de eşantion este considerat a fi cel mai ieftin, cel mai rapid şi mai comod mijloc de eşantionare, care nu necesită nici măcar informații complete despre populație. Se pretează mai degrabă la analize de tip calitativ, cum ar fi interviurile sau focus-grupurile.

Un cluster reprezintă o structură care se comportă ca un individ într-o populație. Dacă considerăm populația municipiilor din România, atunci un cluster este reprezentat de Municipiul Constanța, un alt cluster de Municipiul Iaşi şi aşa mai departe. Modalitatea de realizare a unui eşantion pe clusteri este următoarea:

- **Stabilirea bazei de eşantionare.** În acest caz, baza de eşantionare este reprezentată de totalitatea populației de clustere. Dacă suntem interesați de o cercetare pe învățământul liceal, baza de eşantionare va conține lista exhaustivă a liceelor din România. Să presupunem că avem 5423 de licee în România (*N=5423*);
- **Stabilirea mărimii eşantionului.** Prin procedee specifice de calcul se determină mărimea eşantionului pe care urmează să-l selectăm. În cazul nostru, considerăm un număr de 12 licee (*n=12*);
- **Generarea numerelor aleatoare**. Se generează pe computer 12 de numere aleatoare sau se utilizează tabelul cu numere aleatoare;
- **Extragerea eşantionului**. Se citeşte primul număr aleatoriu generat sau primul număr din tabelul de numere aleatoare. Acel număr reprezintă numărul de ordine al liceului din baza de eşantionare care va fi inclus în eşantion. Se trece apoi la al doilea număr aleatoriu, acesta reprezentând numărul de ordine al celui de-al doilea liceu şi se continuă astfel până la completarea efectivului (12 licee) eşantionului.

Specifică acestui tip de eşantion este desfăşurarea cercetării la nivelul întregului cluster. Într-un liceu selectat în eşantion se investighează toți profesorii, toți elevii, personalul auxiliar.

Această metodă este destul de des utilizată în investigații psihologice sau psiho-sociologice, predominante fiind totuşi cercetările calitative.

VIII.3.4.6 Eşantionarea sistematică

Se numeşte şi **eşantionare cu pas** şi reprezintă o variantă a eşantionării aleatorii simple, fiind chiar mai comodă decât aceasta. Ca principale avantaje menționăm uşurința în utilizare şi posibilitatea de calcul a erorii de eşantionare. Pe lângă dezavantajele eşantionării aleatorii simple, metoda prezintă şi dezavantajul periodicității. Modalitatea de lucru în vederea construirii unui eşantion prin eşantionare stratificată este următoarea:

- **Stabilirea bazei de eşantionare**. Baza de eşantionare se stabileşte prin acelaşi procedeu ca şi în cazul eşantionării aleatorii simple. Avem din nou nevoie de lista exhaustivă a bărbaților din România, spre exemplu;
- **Se determină mărimea eşantionului**. Prin procedee de calcul specifice, stabilim mărimea eşantionului. Pentru exemplul nostru, vom considera un eşantion de 1000 de persoane (*n=1000*);
- **Se stabileşte pasul de eşantionare**. Pasul de eşantionare reprezintă expresia raportului dintre volumul populației şi volumul eşantionului: $pas = {}^{N}/_{n}$. În cazul nostru, pasul de eşantionare va fi ${}^{8000000}/_{1000} = 8000$.
- **Stabilirea punctului de start**. Punctul de start reprezintă locul de începere a eşantionului. Se generează un număr aleatoriu între 1 şi 8000, număr care va reprezenta punctul de start. Să presupunem că acest număr este 5421. Subiectul cu acest număr de ordine va fi prima persoană inclusă în eşantion.
- **Elaborarea eşantionului**. La numărul care indică punctul de start, se adaugă valoarea pasului de eşantionare

(5421+8000=13.421). Subiectul cu acest număr de ordine în baza de eşantionare va fi al doilea subiect inclus în eşantion. Al treilea subiect inclus în eşantion va fi cel cu numărul de ordine 13.421+8000=21.421 şi aşa mai departe până la completarea efectivului de 1000 de persoane.

VIII.3.4.7 Eşantionarea pe cote

Se aseamănă cu eşantionarea stratificată, însă nu mai are un caracter aleatoriu şi de aceea metoda este inclusă în categoria procedeelor de eşantionare nealeatorii. Principul este acelaşi. Se identifică o serie de straturi la nivelul populației şi se încearcă reproducerea acestora în eşantion, fără a se folosi însă tehnici aleatorii. Printre avantajele acestei metode putem menționa rapiditatea obținerii eşantionului, relativa economie de resurse şi lipsa bazei de eşantionare. Totuşi, metoda nefiind aleatorie, riscul de obținere a erorilor de eşantionare este foarte mare. De asemenea, sunt necesare cunoştințe solide referitoare la structura populației iar eroarea de eşantionare şi nivelul de încredere sunt elemente ce nu pot fi calculate. Reluând exemplul din județul Blagosloveşti, am putea construi un eşantion pe cote prin următorul procedeu:

- **Stabilirea cotelor (variabilelor de stratificare) şi a structurii populației.** Se face identic ca şi în cazul eşantionării stratificate. Vor rezulta cele patru grupe, cu procentele din populație asociate (vezi tabelul 8.1);
- **Stabilirea mărimii eşantionului.** Se face în urma unor calcule specifice, adecvate acestui tip de eşantionare, calcule pe care le vom prezenta într-un capitol ulterior.
- **Stabilirea compoziției eşantionului.** Este o etapă în care se operează ca şi la eşantionarea stratificată, stabilindu-se numărul de persoane cuprinse în eşantion, în funcție de proporția în

care se regăsesc în populaţie (vezi tabelul 8.2). Aceasta este şi etapa finală a eşantionării pe cote, nefiind implicată vreo procedură aleatorie sau vreo bază de eşantionare.

Operatorilor li se comunică doar numărul subiecţilor şi categoria din care fac parte, rămânând la latitudinea lor alegerea concretă a cazurilor. Acest lucru induce erori de eşantionare serioase, determinând includerea acestei metode în categoria procedeelor nealeatorii. De exemplu, dacă avem un număr de 10 operatori pentru realizarea studiului de mai sus, acestora li se va comunica doar necesitatea administrării instrumentelor de cercetare pe un număr de 8 bărbaţi din mediul rural, 5 bărbaţi din mediul urban, 8 femei din mediul rural şi 9 femei din mediul urban, fără vreo referire la persoane concrete.

Eşantioanele pe cote se aplică destul de des în procesul de adaptare a testelor psihologice în România, instrumentele rezultate fiind şi ele la rândul lor limitate de problemele acestei metode.

VIII.3.4.8 Eşantioane cu probabilitatea proporţională cu mărimea

Reprezintă o metodă derivată din eşantionarea sistematică, prin intermediul căreia probabilitatea de selectare a unei unităţi de eşantionare este proporţională cu dimensiunea acesteia.

Tehnica este extrem de utilă în condiţiile în care unităţile de eşantionare diferă foarte mult sub aspectul dimensiunilor, determinând ca unităţile de eşantionare mari să aibă aceeaşi probabilitate de a fi selectate ca şi cele mici. Între Bucureşti şi Paşcani există o mare diferenţă în ceea ce priveşte numărul de locuitori. În mod normal, probabilitatea de a selecta într-un eşantion mai mulţi bucureşteni este mult mai mare. Folosind această tehnică, probabilităţile se egalează. Eşantioanele cu probabilitate proporţională cu mărimea sunt utilizate, în special, în domeniul organizaţional, în cercetări la nive-

lul cărora se intenționează investigarea personalului. Deoarece cadrele de conducere sunt mai puțin numeroase în comparație cu cadrele de execuție, metoda permite egalarea şanselor celor două categorii de a fi incluse în eşantion.

Reluând exemplul anterior, ne propunem să desfăşurăm o cercetare oarecare în județul Trubadurul, care are un număr de 29 de localități. După cum observați în tabelul 8.3, localitățile sunt diferențiate sub aspectul populației. În timp ce localitatea 26 are un număr de 197 de locuitori, localitatea 21 are un număr de 1590 de locuitori. Evident că şansele de a extrage mai mulți locuitori din localitatea 21 sunt mai mari în comparație cu localitatea 26, de unde, cel mai probabil, nu va participa nimeni la cercetare. Prin acest procedeu de eşantionare putem egala situația, astfel încât subiecții din fiecare localitate să aibă o şansă calculabilă de a fi incluşi în eşantion.

Etapele de realizare a unui asemenea eşantion sunt următoarele:

- **Inventarierea unităților de eşantionare şi a dimensiunii acestora.** În această etapă se listează toate unitățile de eşantionare din teritoriu, precum şi volumul corespunzător de populație. În cazul nostru, vom înregistra toate cele 29 de localități – ordinea nu contează – împreună cu volumul populației pentru fiecare unitate (coloanele Localitate şi Populație din tabelul 8.3). La acest nivel se poate lucra fie cu lista completă, fie cu unități de eşantionare grupate în interval (de exemplu vârsta între 0 şi 10 ani, între 11 şi 21 de ani etc.)

Tabel 8.3 – Eşantion cu probabilitate proporțională cu mărimea

Localitate	Populație	Populație cumulată	Locație selectată
1	542	542	423
2	245	787	
3	1032	1819	
4	867	2686	1891
5	256	2942	
6	352	3294	

7	835	4129	3359
8	645	4774	
9	427	5201	4827
10	312	5513	
11	1342	6855	6295
12	390	7245	
13	604	7849	7763
14	465	8314	
15	897	9211	
16	476	9687	9231
17	365	10052	
18	967	11019	10699
19	533	11552	
20	215	11767	
21	1590	13357	12167
22	423	13780	13635
23	645	14425	
24	867	15292	15103
25	423	15715	
26	197	15912	
27	586	16498	
28	365	16863	16571
29	756	17619	

- **Calculul populaţiei cumulate.** Vom calcula apoi frecvenţa cumulată a întregului volum de populaţie ţintă. Ultima unitate de eşantionare va conţine, evident, volumul total al populaţiei. În cazul nostru, în judeţul Trubadurul trăiesc 17.619 persoane (*N=17.619*). Ne reamintim că prin calculul frecvenţelor cumulate, practic „ordonăm" crescător populaţia noastră.

- **Stabilirea mărimii eşantionului.** Prin procedee specifice de calcul se determină mărimea eşantionului. În cazul nostru vom considera că am ales un eşantion de 120 de persoane, reprezentativ pentru judeţul Trubadurul (*n=120*).

- **Determinarea numărului de locaţii.** Eşantionul de 120 de persoane va trebui ales din unele dintre cele 29 de localităţi ale judeţului Trubadurul. În funcţie de numărul total al unităţilor

de eşantionare, se stabilesc localităţile în care se va desfăşura cercetarea. Pentru 120 de persoane, am ales un număr de 12 localităţi din cele 29. În fiecare localitate vor fi intervievate 10 persoane.

- **Stabilirea pasului de eşantionare.** Se realizează la fel ca la eşantionarea sistematică, prin divizarea populaţiei totale, de data aceasta la numărul locaţiilor şi nu la volumul eşantionului. Pentru exemplul nostru, avem un pas de eşantionare de 17.619:12=1468,25 rotunjit 1468.

- **Stabilirea punctului de start.** Se va alege aleatoriu un număr între 1 şi pasul de eşantionare, acela fiind numărul locaţiei de start. Să presupunem că numărul generat aleatoriu este 423. La fel ca la stabilirea medianei, verificăm categoria ce conţine al 423-lea subiect, prin inspectarea coloanei „Populaţie cumulată". Observăm că această valoare cumulată se regăseşte chiar în prima locaţie. Prin urmare, una dintre cele 12 locaţii este localitatea 1. În acea localitate, folosindu-se ca bază de eşantionare lista exhaustivă a celor 542 de persoane, ori prin altă metodă aleatorie sau sistematică (de exemplu de la casa 10, fiecare a 54-a casă), vor fi intervievate 10 persoane.

- **Stabilirea eşantionului.** La punctul de start 423 se adaugă valoarea pasului 1468 şi se obţine noua locaţie (423+1468=1891). Valoarea cumulată pentru a doua locaţie se află în dreptul localităţii 4 (Vă mai amintiţi, sper, că la nivelul localităţii 3 se află a 1819-a persoană. Persoana cu numărul 1891 se află în localitatea 4, care cuprinde subiecţii de la al 1820-lea la al 2686-lea). Un alt operator se va deplasa în această localitate, şi printr-un procedeu similar va intervieva

10 persoane. Procesul continuă până la stabilirea tuturor celor 12 localităţi.

Au rezultat exact 12 localităţi, mai precis localităţile 1 – 542 locuitori, 4 – 867 locuitori, 7 – 835 locuitori, 9 – 427 locuitori, 11 – 1342 locuitori, 13 – 604 locuitori, 16 – 476 locuitori, 18 – 967 locuitori, 21 – 1590 locuitori, 22 – 423 locuitori, 24 – 867 locuitori şi 28 – 365 locuitori.

VIII.3.5 Determinarea mărimii eşantionului

În toate exemplele din subcapitolul anterior am stabilit valori exacte pentru mărimea unui eşantion. Totuşi, întrebarea pe care cred că v-o puneţi este cum determinăm mărimea eşantionului, astfel încât folosind o anumită tehnică de eşantionare, eşantionul rezultat să fie reprezentativ. Cel mai simplu răspuns ar fi să consultaţi pagina de internet „The Survey System", aflată la adresa: http://www.surveysystem.com/sscalc.htm unde găsiţi un excelent calculator pentru mărimea eşantionului, împreună cu explicaţiile de rigoare. Un asemenea demers nu ar fi însă ştiinţific, de aceea vom descoperi împreună procedeele necesare stabilirii mărimii eşantionului.

Determinarea mărimii unui eşantion depinde, în principal, de cinci factori:

- **Eroarea statistică –** ca expresie a diferenţei dintre indicatorul obţinut la nivelul eşantionului şi valoarea parametrului la nivelul populaţiei. Am mai spus acest lucru; prin studiul unui eşantion noi nu putem decât să estimăm un parametru al populaţiei. Niciodată nu obţinem valoarea sa exactă. În funcţie de modul în care proiectăm cercetarea şi de nivelul de precizie dorit, putem reduce eroarea statistică, însă în condiţiile unui eşantion de dimensiuni mai mari;
- **Variabilitatea populaţiei –** estimată de obicei prin intermediul abaterii standard. Se referă la amplitudinea pe care o au

opiniile populației, element care afectează de asemenea dimensiunea eşantionului. Cu cât variabilitatea este mai mare, cu atât acuratețea estimării este mai mică şi va fi nevoie de un eşantion de dimensiuni mai mari;

- **Nivelul de încredere –** indică gradul de precizie cu care rezultatele obținute prin studiul eşantionului se vor regăsi la nivelul populației. Cu cât nivelul de încredere este mai mare, cu atât pretenția că rezultatele obținute nu sunt atipice creşte, fiind nevoie de eşantioane de dimensiuni mai mari. Majoritatea cercetărilor utilizează un nivel de încredere de 95%;
- **Dimensiunea populației** – în mod normal, acest factor nu afectează dimensiunea eşantionului atâta timp cât aceasta se află sub 5% din populație. Dacă dimensiunea eşantionului este mică în raport cu dimensiunea populației, factorul se poate ignora. Dimensiunea populației devine importantă în condițiile în care mărimea eşantionului depăşeşte 5% din totalul populației (de exemplu, dacă extragem un eşantion de 40 de persoane dintr-o populație de 100 de persoane). În general, efectul este acela de a se reduce uşor dimensiunea eşantionului.
- **Proporția din populație** – se referă doar la situația în care eşantionul este folosit pentru a măsura un atribut din populație în locul unei variabile continue. Spre exemplu, opinia populației referitoare la realegerea preşedintelui, ori la efectele nocive ale fumatului. Subiecții se pot exprima prin „Da" sau „Nu", şi atunci ne interesează proporția opiniilor pozitive şi a celor negative. Proporția estimată a acestor răspunsuri influențează dimensiunea eşantionului.

În mod curent, cercetările ştiințifice pe eşantioane de 50 – 100 de persoane sunt suficient de precise în domeniul socio-uman. Poate doar sociolo-

gia presupune eşantioane mai mari, în cazul sondajelor de opinie. Uneori este suficient un eşantion de 30 de subiecți, considerat ca fiind volumul minim de date ce permite analize statistice, respectiv aplicarea legii numerelor mari şi a distribuției gaussiene. Eşantioanele cu dimensiuni sub 30 de cazuri fac parte din domeniul studiilor de caz, rezultatele obținute neputând fi extinse la nivelul populației.

VIII.3.5.1 Calculul mărimii eşantionului

Mărimea eşantionului de calculează în funcție tipul variabilei de interes. Dacă variabila de interes măsoară un **atribut**, atunci stabilirea mărimii eşantionului se face prin intermediul proporțiilor; dacă este o **variabilă continuă**, mărimea eşantionului se determină pe baza valorilor sau se converteşte variabila continuă în atribut dihotomic.

Mărimea eşantionului pentru o variabilă care măsoară un atribut (întrebări de tip „Da" şi „Nu" de exemplu) se determină pe baza formulei lui Cochran:

$$n = \frac{z^2(p)(1-p)}{e^2} \qquad \textit{(formula 8.1)}$$

unde: n reprezintă mărimea eşantionului, z – scorul „z" asociat nivelului de încredere dorit, e – este precizia dorită (numită şi marjă de eroare) şi p – reprezintă ponderea răspunsurilor la nivelul populației.

După cum v-ați dat seama, valorile notei „z" sunt de fapt nişte constante, alocate în funcție de nivelul de încredere, astfel:

- La un nivel de încredere de 80%, nota „z" este 1,28
- La un nivel de încredere de 85%, nota „z" este 1,44
- La un nivel de încredere de 90%, nota „z" este 1,65
- La un nivel de încredere de 95%, nota „z" este 1,96

- La un nivel de încredere de 99%, nota „z” este 2,58

Deşi scorurile pentru notele „z” au fost prezentate de la un nivel de încredere de 80%, ştim foarte bine că nivelul de încredere minim pentru o cercetare serioasă este de 95%. De aceea, nu recomand niciodată să utilizați alte scoruri „z” în afară de 1,96 şi 2,58.

Ponderea răspunsurilor la nivelul populației reprezintă un estimator a valorii aşteptate pe care o poate lua atributul respectiv atunci când ne referim la populația țintă. Situația fericită este când anticipăm diferențe procentuale mari. Spre exemplu, când putem bănui că 95% din populație ar prefera intrarea în Uniunea Europeană, în timp ce 5% din populație nu ar prefera-o, suntem în fața unei **soluții optimiste**, cu o pondere a răspunsurilor de 0,05 sau de 0,95 – în funcție de sensul în care ne orientăm. În acest caz, volumul eşantionului va fi mic (75 de persoane la un nivel de încredere de 95%). **Situația pesimistă** este aceea în care nu putem anticipa orientarea atributului. În acest caz, suntem nevoiți să presupunem o pondere de 0,5, adică jumătate din populație ar răspunde „Da”, iar cealaltă jumătate „Nu”. Volumul eşantionului va fi acum semnificativ mai mare (384 de subiecți la un nivel de încredere de 95%).

Precizia se referă la limitele marjei de eroare între care acceptăm să se situeze valoarea indicatorului, exprimată sub formă zecimală. O marjă de eroare ±2% la un procent de, să spunem, 53% răspunsuri „Da”, obținut prin studiul eşantionului, înseamnă că, la nivelul populației, procentul de răspunsuri „Da” se situează undeva între 51% şi 55%. Desigur, cu cât această precizie este mai mare, cu atât creşte şi volumul eşantionului. Marja de eroare de ±2% se scrie în formă zecimală 0,02.

Cunoscând acum toate aceste elemente, să ne imaginăm că desfăşurăm un studiu utilizând un chestionar, la care subiecții răspund prin „Da” sau „Nu”, nu avem idee care ar fi proporția răspunsurilor „Da” sau care ar fi pro-

porția răspunsurilor „Nu”, dorim un nivel de încredere de 95% şi un nivel de precizie de ±2%. Ce volum trebuie să aibă eşantionul?

$$n = \frac{z^2(p)(1-p)}{e^2} = \frac{1{,}96^2(0{,}5)(1-0{,}5)}{0{,}02^2} = \frac{3{,}84\times0{,}5\times0{,}5}{0{,}0004} = \frac{0{,}96}{0{,}0004} = 2400$$

Vom avea un eşantion de 2400 de persoane pentru a obține reprezentativitatea la acest nivel. Suntem cam pretenţioşi, nu-i aşa? Bine că nu am ales nivelul de încredere de 99%, că cine ştie ce obțineam… Dacă doriți să aflați, înlocuiți în formula de mai sus 1,96 cu 2,58 şi vedeți.

Exercițiu:

Folosind acelaşi tip de studiu, calculați volumul unui eşantion la un nivel de încredere de 95% şi cu o precizie de ±3%, estimând că subiecții vor răspunde 75% „Da” şi 25% „Nu”.

În practică se foloseşte o formulă derivată din formula 8.1 în vederea calculării preciziei sau a marjei de eroare:

$$e = \sqrt{\frac{z^2(p)(1-p)}{n}} \qquad \text{(formula 8.2)}$$

În presa scrisă, televiziune, la radio, auziți frecvent fraze de acest gen: „65% din populație îl preferă pe Icsulescu la alegerile prezidențiale. Sondajul are o marjă de eroare de ±2,5%.” Cunoscând marja de eroare şi faptul că nivelul de încredere nu se poate situa sub 95%, puteți afla uşor din câte persoane a fost format eşantionul, aplicând formula 8.1. Dar dacă afirmația de mai sus nu prezintă marja de eroare, ci doar volumul eşantionului? Dacă vi se comunică rezultatele de mai sus, cu specificația că eşantionul a fost format dintr-un număr de 180 de persoane, cum aflați precizia? Eu aş fi de părere să aplicăm formula 8.2

$$e = \sqrt{\frac{z^2(p)(1-p)}{n}} = \sqrt{\frac{1{,}96^2 \times (0{,}65) \times (1-0{,}65)}{180}} = \sqrt{\frac{3{,}84 \times 0{,}65 \times 0{,}35}{180}} = \sqrt{\frac{0{,}87}{180}} = \sqrt{0{,}004} = 0{,}06 = 6\%$$

O precizie de ±6% mi se pare extrem de mică pentru un asemenea sondaj. Cu alte cuvinte, preferințele pentru Icsulescu variază între 59% şi 71%. Cam mare marja de eroare, nu credeți?

Atunci când variabila măsurată are un caracter continuu (cum ar fi de exemplu înălțimea), se utilizează o altă formulă pentru determinarea mărimii eşantionului, formulă ce ține cont de varianța caracteristicii măsurate la nivelul populației. Deoarece această varianță de multe ori nu poate fi anticipată, existând prea puține şanse să cunoaştem în mod real varianța parametrului la nivelul populației, majoritatea studiilor de acest tip preferă transformarea variabilelor continui sau polihotomice în variabile dihotomice pentru aplicarea calculului de proporții. Acest lucru se poate efectua prin combinarea răspunsurilor în două categorii (de exemplu subiecți înalți şi subiecți scunzi) şi apoi aplicarea metodei de estimare prin proporții, această tehnică fiind acceptată cvasi-universal.

Dacă nu dorim totuşi să facem acest lucru, o altă metodă prin care putem determina volumul unui eşantion în cazul variabilelor continui este aceea prin care înlocuim varianța populației cu varianța eşantionului. Deoarece abaterea standard la nivelul unui eşantion aproximează bine abaterea standard a populației, ne putem permite să realizăm un studiu pilot pe un număr de 30-40 de subiecți, calculăm apoi abaterea standard, pe care o folosim la determinarea mărimii eşantionului după formula:

$$n = \frac{z^2 \times \sigma}{e^2} \qquad \text{(formula 8.3)}$$

Spre exemplu, dorim să efectuăm o cercetare în care să putem stabili înălțimea medie a bărbaților din România. Dacă am şti abaterea standard a parametrului „înălțime" pentru întreaga populație, ar fi perfect. O asemenea

valoare probabil că nu poate fi niciodată cunoscută. Totuşi avem nevoie să stabilim volumul unui eşantion reprezentativ de bărbaţi, la un nivel de încredere de 95% şi cu o marjă de eroare de 0,2. Nu cunoaştem însă valoarea abaterii standard. Pentru a o afla, vom stabili un lot de 30 de bărbaţi, aleşi întâmplător, şi vom măsura înălţimea acestora. Datele au fost deja analizate într-un capitol anterior, de unde am reţinut că abaterea standard avea valoarea 4,46. Nu ne rămâne acum decât să înlocuim în formulă şi obţinem:

$$n = \frac{1{,}96^2 \times 4{,}46}{0{,}2^2} = 428$$

Formulele de mai sus sunt aplicabile numai în condiţia în care volumul eşantionului nu depăşeşte 5% din volumul populaţiei. În situaţia în care volumul eşantionului trece de acest prag, se aplică o corecţie care-i reduce uşor dimensiunea, după formula:

$$n = \frac{n}{1+\frac{n-1}{N}} \qquad \text{(formula 8.4)}$$

Dacă într-o cercetare am stabilit volumul unui eşantion la 398 de cazuri iar volumul populaţiei (N) este de 2000 de cazuri, aplicând formula vom obţine:

$$n = \frac{n}{1+\frac{n-1}{N}} = \frac{398}{1+\frac{398-1}{2000}} = \frac{398}{1{,}19} = 334$$

S-a constatat o uşoară reducere a eşantionului către pragul de 5%. Această metodă poartă numele de **corecţie pentru populaţii finite**. Yamane a stabilit o formulă directă de calcul a eşantionului pentru populaţii de mici dimensiuni, pe baza formulei:

$$n = \frac{N}{1+N(e)^2} \qquad \text{(formula 8.5)}$$

Dacă vom aplica formula pe cazul anterior, vom obține la un nivel de precizie de ±5%:

$$n = \frac{N}{1 + N(e)^2} = \frac{2000}{1 + 2000 \times 0{,}05^2} = \frac{2000}{6} = 333$$

Observăm că valoarea se apropie mult de cea obținută anterior prin corecție.

VIII.4 Construcția eşantioanelor în SPSS

SPSS prezintă un instrument foarte puternic de creare a eşantioanelor în condițiile în care aveți la dispoziție baza de eşantionare. Cu ajutorul programului SPSS puteți crea următoarele tipuri de eşantioane:

- Eşantioane simple aleatorii, cu sau fără înlocuire;
- Eşantioane stratificate şi pe clusteri;
- Eşantioane multistadiale;
- Eşantioane probabilistice proporțional cu mărimea sau tipuri de eşantioane nealeatorii.

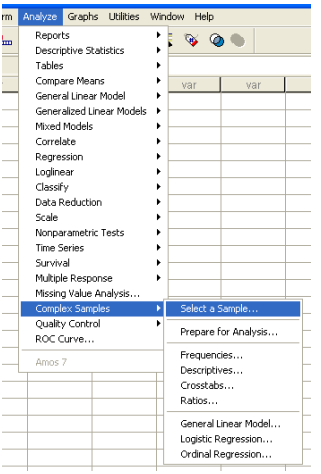


Figura 8.4 – Meniul de creare şi analiză a eşantioanelor

Structura eşantioanelor în SPSS se prezintă sub forma unor fişiere tip plan. Există un număr de două planuri:

- **Planul de eşantionare** – conține specificațiile de definire a eşantionului.
- **Planul de analiză** – conține date necesare procedurilor de analiză complexă pentru a calcula varianța eşantionului. Planul

include structura eşantionului, metodele de estimare pentru fiecare stadiu, referinţele la variabilele de interes.

Debutul procedurii de creare a unui eşantion se face prin deschiderea bazei de eşantionare. Vom folosi, în acest scop, o bază de date pusă la dispoziţie de SPSS Inc. situată în directorul în care s-a instalat aplicaţia. Baza de date se numeşte „*1991 U.S. General Social Survey.sav*" şi conţine un număr de 1517 înregistrări reprezentând rezultatele unui sondaj efectuat în Statele Unite. Deoarece volumul de date este destul de mare, având şi o serie de variabile factuale şi demografice, putem folosi acest fişier ca exemplu pentru baza noastră de eşantionare.

VIII.4.1 Crearea unui eşantion în SPSS

Proiectarea unui nou plan de eşantionare se realizează prin apelarea meniului **„Analyze"** iar din submeniul **„Complex Samples"** se alege opţiunea **„Select a Sample…"**. Crearea unui eşantion este asistată de un program special („Wizard" sau asistent) ce presupune mai mulţi paşi.

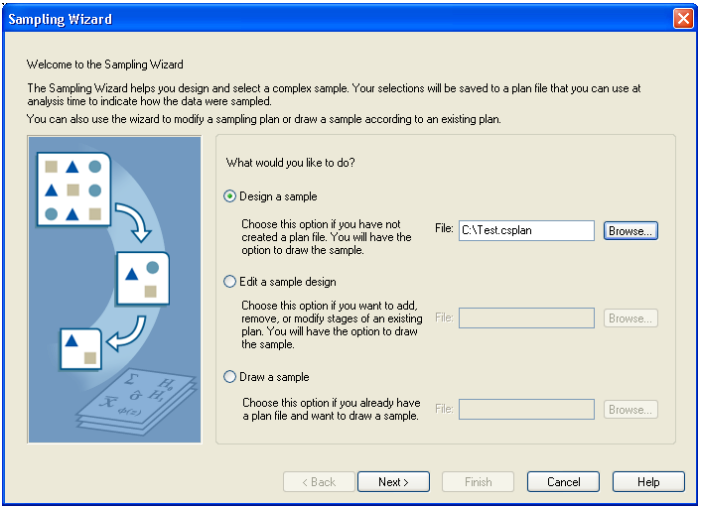

Sampling Wizard
Welcome to the Sampling Wizard
The Sampling Wizard helps you design and select a complex sample. Your selections will be saved to a plan file that you can use at analysis time to indicate how the data were sampled.
You can also use the wizard to modify a sampling plan or draw a sample according to an existing plan.
What would you like to do?
Design a sample
Choose this option if you have not created a plan file. You will have the option to draw the sample.
File: C:\Test.csplan
Browse...
Edit a sample design
Choose this option if you want to add, remove, or modify stages of an existing plan. You will have the option to draw the sample.
File:
Browse...
Draw a sample
Choose this option if you already have a plan file and want to draw a sample.
File:
Browse...
< Back
Next >
Finish
Cancel
Help

Figura 8.5 – Primul pas al creării unui eşantion. Salvarea fişierului plan

Prima fereastră afişată ne solicită alegerea unui fişier în care va fi stocat planul de eşantionare. Opțiunea implicită este **„Design a sample”** pe care o vom folosi pentru a specifica locația şi numele planului de eşantionare. Vom putea introduce direct calea către fişierul ce urmează a fi salvat sau vom putea apăsa butonul **„Browse…”** pentru a alege locul pe hard-disc unde va fi salvat planul de eşantionare. Evident, va trebui să furnizăm şi un nume pentru acest fişier. În cazul nostru, fişierul l-am denumit *„Test.csplan”* şi a fost salvat direct pe discul *C:*. Observați că extensia fişierelor de tip plan de eşantionare este *„.csplan”*.

Pentru a modifica un plan de eşantionare existent, aveți la dispoziție opțiunea **„Edit a sample design”**. Cu ajutorul acesteia veți putea reconfigura sau elimina stadii din cadrul unui plan de eşantionare multistadial creat anterior. Evident, acest plan de eşantionare va trebui inițial deschis prin tastarea căii către fişierul de tip *.csplan* sau folosind butonul **„Browse…”**.

Dacă aveți deja un plan de eşantionare făcut şi doriți să creați un eşantion în baza acestuia, puteți folosi opțiunea **„Draw a sample”**, după care alegeți fişierul plan dorit, la fel ca mai sus. Trecerea la următoarea etapă se face prin intermediul butonului **„Next”.** Puteți oricând abandona operația apăsând pe butonul **„Cancel”** sau puteți parcurge sistemul de asistență contextuală prin accesarea butonului **„Help”.**

A doua fereastra presupune definirea planului de eşantionare. În partea stângă, având o structură arborescentă, sunt afişate opțiunile acestui stadiu. Prima opțiune implică precizarea variabilelor de eşantionare în cazul unor eşantioane ce presupun acest lucru (opțiunea **„Design variables”**). Avem prezentate, în secțiunea centrală, toate variabilele din baza de eşantionare. Folosind butoanele de transfer (butoanele săgeată) vom putea include aceste variabile în secțiunea **„Stratify By”** dacă dorim un eşantion stratificat după acea variabilă. Se vor putea preciza astfel straturile sau subpopulațiile (amintiți-vă de fasolele de 1 an şi de 10 ani). În urma acestei selecții se obțin

eşantioane separate pentru fiecare strat. Pentru a se putea îmbunătăți precizia estimărilor, unitățile din fiecare strat vor fi cât mai omogene din punctul de vedere al caracteristicii estimate.

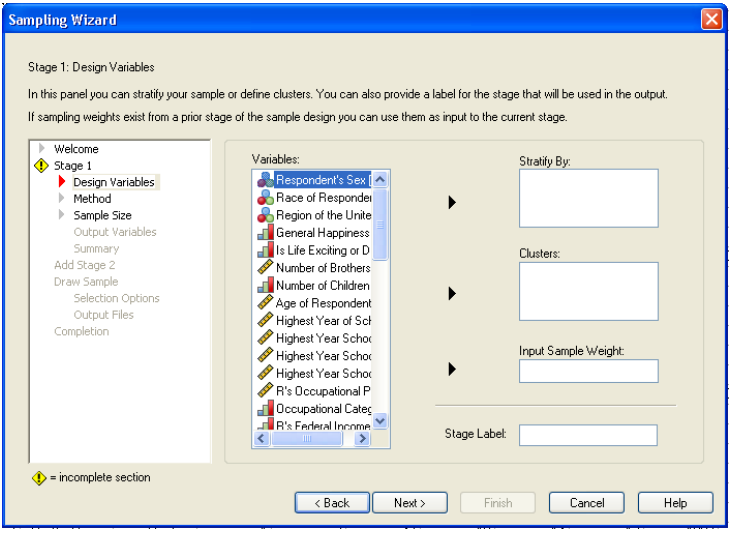


Figura 8.6 – Pasul 2: Definirea variabilelor cercetării

O variabilă introdusă în secțiunea **„Clusters”** permite crearea unui eşantion pe clusteri în care se definesc grupuri de unități de observație. Pentru a se evita pierderea preciziei şi a compensa dezavantajele eşantionării pe clusteri, între clusteri trebuie să existe o eterogenitate cât mai mare sub aspectul variabilei de interes.

Dacă planul de eşantionare este parte dintr-un plan de eşantionare mai complex, puteți introduce o variabilă numerică în secțiunea **„Input Sample Weight”**, variabilă ce reprezintă dimensiunea unui eşantion construit într-un stadiu anterior.

Pentru claritate, se poate introduce un scurt text care va descrie stadiul în curs, folosind caseta de text **„Stage Label”**.

În acest exemplu vom dori să construim un eşantion simplu randomizat. Deocamdată nu ne interesează în mod deosebit opţiunile prezentate.

Următoarea secţiune, **„Method”**, permite stabilirea tehnicii de eşantionare pe care o folosim.

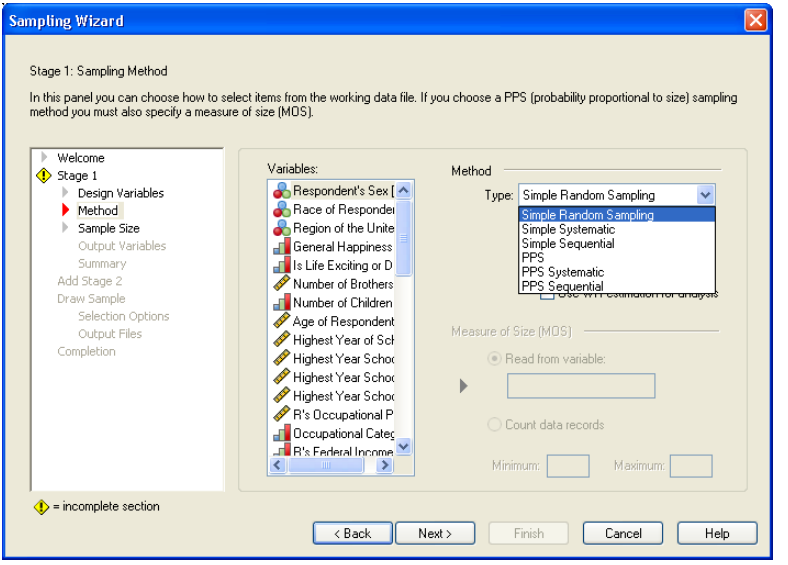


Figura 8.7 – Pasul 2: Alegerea metodei de eşantionare

SPSS permite selectarea, în funcţie de variabilele precizate în etapa anterioară, a mai multor tipuri de eşantioane. Deoarece nu am inclus nici o variabilă în etapa anterioară – optând pentru un eşantion simplu randomizat – în caseta de selecţie **„Type”** din cadrul secţiunii **„Method”** vor fi disponibile doar opţiunile de acest tip. Observăm existenţa doar a eşantioanelor de tip randomizat – simplu, sistematic şi secvenţial şi a eşantioanelor de tip proporţional cu mărimea – PPS simplu, sistematic şi secvenţial. Din această listă vom alege eşantionul simplu randomizat (opţiunea **„Simple Random Sample”**).

La nivelul acestei etape există posibilitatea de a efectua selecția fără „reintroducerea bilei în urnă" – opțiunea **„Without replacement (WOR)"** sau „cu reintroducerea bilei în urnă" – opțiunea **„With replacement (WR)"**. Cu alte cuvinte, o unitate selectată în eşantion va fi eliminată de la următoarele posibile selecții – în cazul WOR – sau nu va fi eliminată, existând posibilitatea unei noi selecții – în cazul WR. În acest din urmă caz, bifând caseta **„Use WR estimation for analysis"** comunicați programului SPSS să estimeze această probabilitate de reincludere a unui element în eşantion. Implicit, metoda de estimare este specificată în fişierul plan şi este consistentă cu tipul de eşantion selectat. Opțiunea permite, de asemenea, forțarea metodei cu introducerea bilei în urnă, chiar dacă metoda de eşantionare presupune o selecție fără introducerea bilei în urnă.

Pe scurt, tipurile de metode ce pot fi alese în această etapă sunt:

- **Simple Random Sampling** – eşantion aleatoriu simplu în care unitățile sunt selectate cu o probabilitate egală. Metoda poate fi folosită cu sau fără reintroducerea bilei în urnă;
- **Simple Systematic** – eşantion sistematizat simplu (cu pas). Unitățile sunt selectate cu un pas fix de eşantionare din baza de eşantionare. Se poate folosi şi în cazul eşantionării stratificate dacă s-au specificat anterior variabilele de stratificare. Selecția se face fără introducerea bilei în urnă;
- **Simple Sequential –** unitățile sunt extrase secvențial cu probabilitate egală şi fără introducerea bilei în urnă;
- **PPS –** eşantion proporțional cu mărimea în care unitățile sunt extrase aleatoriu cu o probabilitate proporțională cu mărimea. Orice unitate se selectează cu reintroducerea bilei în urnă, exceptând configurarea anterioară a unui eşantion pe clusteri;

- **PPS Systematic –** eşantion proporțional cu mărimea de tip sistematic.
- **PPS Sequential –** eşantion proporțional cu mărimea de tip secvențial;
- **PPS Brewer –** eşantion proporțional cu mărimea în care sunt selectate două clustere din fiecare strat cu o probabilitate proporțională cu mărimea clusterului, fără introducerea bilei în urnă. Necesită precizarea unei variabile de tip cluster;
- **PPS Murthy –** un eşantion similar celui anterior. Diferă doar metoda statistică prin care sunt alese unitățile;
- **PPS Sampford –** eşantion proporțional cu mărimea ce selectează mai mult de doi clusteri din fiecare strat cu o probabilitate proporțională cu mărimea clusterului şi fără introducerea bilei în urnă. Necesită, de asemenea, specificarea unui variabile pentru cluster

Opțiunea **„Measure of Size (MOS)"** se aplică eşantioanelor de tip PPS – proporționale cu mărimea – şi implică precizarea unei dimensiuni definitorii pentru mărimea fiecărei unități. Aceste valori pot fi definite în mod explicit, extrase fiind dintr-o variabilă (prin bifarea opțiunii **„Read from variable"** şi introducerea variabilei din care vor fi extrase datele) sau pot fi calculate automat selectând **„Count data records"**. De asemenea se poate preciza amplitudinea acestor dimensiuni introducând valorile minime şi maxime în casetele de text **„Minimum"** respectiv **„Maximum"**.

Determinarea mărimii eşantionului reprezintă ultima acțiune din acest stadiu. Pentru baza noastră de date, un eşantion cu o marjă de eroare de 2% şi un interval de încredere de 95% implică un număr de 930 de subiecți pentru a fi reprezentativ.

Acest pas presupune specificarea numărului sau a proporţiei în care vor fi selectate unităţile de eşantionare. Dimensiunea eşantionului poate fi fixă sau poate varia în funcţie de straturi.

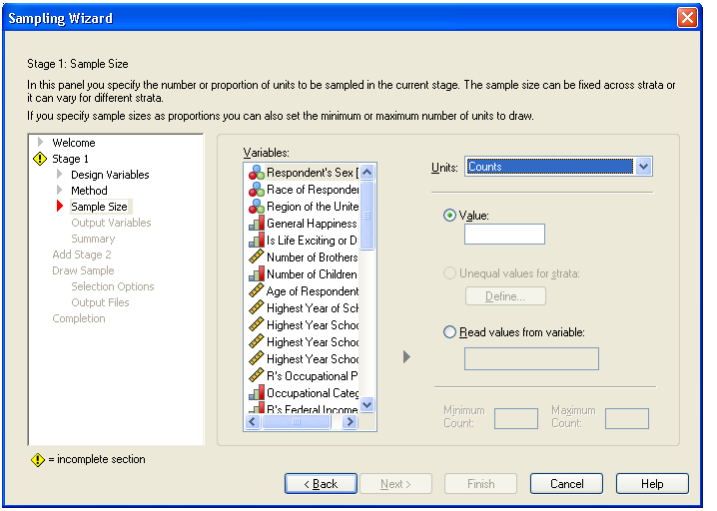


Figura 8.8 – Pasul 2: Stabilirea dimensiunii eşantionului

Caseta de selecţie **„Units”** permite alegerea între un număr fix de unităţi selectate – alegând opţiunea **„Counts”** – sau a unei proporţii de unităţi în eşantion – alegând opţiunea **„Proportions”**.

Opţiunea **„Value”** permite aplicarea unei valori unice pentru toate straturile. În modul numărului fix, va trebui să introduceţi valoarea dimensiunii eşantionului (în cazul nostru 930). Dacă optăm pentru proporţii vom include proporţia din populaţie pe care o regăsim în eşantion (de exemplu 0,10 pentru 10% din populaţie).

Opţiunea **„Unequal values for strata”** permite introducerea valorilor eşantionului pentru fiecare strat definit, în condiţiile în care dimensiunile straturilor nu sunt egale.

„Read values from variable” permite selectarea unei variabile numerice în care sunt stocate dimensiunile fiecărui strat.

La alegerea eşantionului bazat pe proporţii, putem de asemenea specifica limitele inferioare şi superioare ale numărului de unităţi ce vor fi selectate (de exemplu nu mai puţin de 50 – „Minimul” şi nu mai mult de 200 – „Maximum”)

Apăsarea butonului **„Next”** determină afişarea ferestrei variabilelor ce urmează a fi calculate.

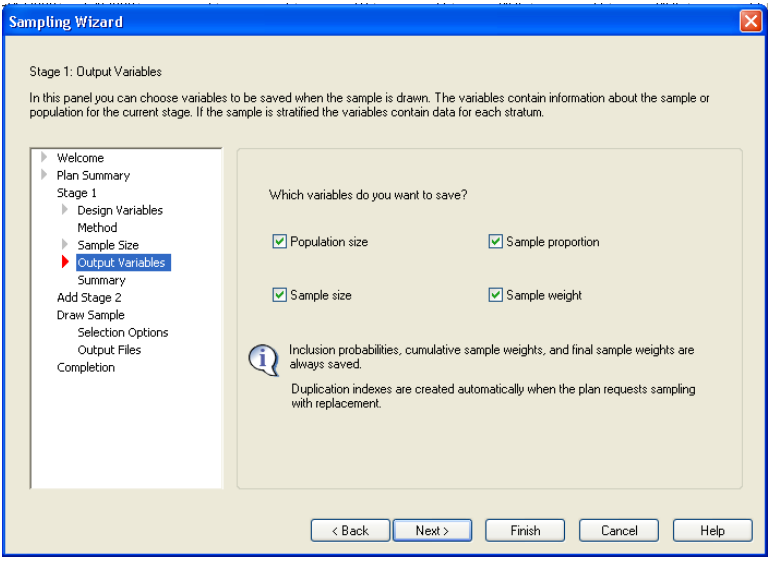

Sampling Wizard
Stage 1: Output Variables
In this panel you can choose variables to be saved when the sample is drawn. The variables contain information about the sample or population for the current stage. If the sample is stratified the variables contain data for each stratum.
Welcome
Plan Summary
Stage 1
Design Variables
Method
Sample Size
Output Variables
Summary
Add Stage 2
Draw Sample
Selection Options
Output Files
Completion
Which variables do you want to save?
Population size
Sample proportion
Sample size
Sample weight
Inclusion probabilities, cumulative sample weights, and final sample weights are always saved.
Duplication indexes are created automatically when the plan requests sampling with replacement.
< Back
Next >
Finish
Cancel
Help

Figura 8.9 – Pasul 3: Alegerea variabilelor eşantionului

Pe parcursul acestei etape putem preciza ce variabile vor fi salvate în baza de date în urma creării eşantionului, astfel:

- Bifând caseta **„Population size”** vom comunica aplicaţiei crearea unei variabile ce va stoca numărul estimat de unităţi din populaţie la un anumit stadiu. În baza de date această variabilă va începe cu *„PopulationSize_”;*

- Bifând caseta **„Sample size”** vom putea salva o variabilă care să conțină numărul de unități selectate la un anumit stadiu. În baza de date această variabilă va începe cu *„SampleSize_”;*
- Bifând caseta **„Sample proportion”** se va crea o variabilă ce va conține proporția eşantionului selectat din populație. Variabila va începe cu *„SamplingRate_”;*
- Bifând caseta **„Sample weight”** comunicăm programului SPSS să creeze o variabilă ce va stoca inversul probabilităților de includere a unui element în eşantion. Numele variabilei va începe în baza de date cu *„SampleWeightCumulative_”* iar variante ale acestei variabile au un rol important în procesul de analiză a eşantionului.

Alături de aceste variabile, în baza de date vor fi create automat o serie de noi variabile după cum urmează:

- Probabilitatea de includere – proporția unităților selectate la nivelul unui anumit stadiu, variabile identificată prin *„InclusionProbability_”;*
- Mărimea cumulată – mărimea cumulată a eşantionului de la un stadiu la altul, variabilă identificată prin *„SampleWeightCumulative_”;*
- Indexul – permite specificarea unităților selectate de mai multe ori la nivelul unui stadiu, variabilă identificată prin *„Index_”.* Variabila este salvată doar în condițiile în care se foloseşte opțiunea „cu introducerea bilei în urnă”.

Ca regulă generală, în baza de date fiecare variabilă nou creată va fi urmată, după linia de subliniere, de o cifră care indică stadiul. De exemplu,

probabilitatea de includere a unei unități în eşantion pentru primul stadiu va apărea în baza de date sub forma *„InclusionProbability_1_”*.

În exemplul nostru, am bifat salvarea tuturor variabilelor în baza de date în vederea comentării ulterioare a câtorva cazuri. Apăsarea butonului **„Next”** permite trecerea la următorul pas al creării eşantionului. Am observat că la un moment dat s-a activat şi butonul **„Finish”**. Acest lucru înseamnă că SPSS are deja suficiente informații pentru a crea eşantionul şi putem sări peste paşii rămaşi. Noi vom parcurge totuşi întregul proces pentru a putea analiza toate etapele acestui instrument.

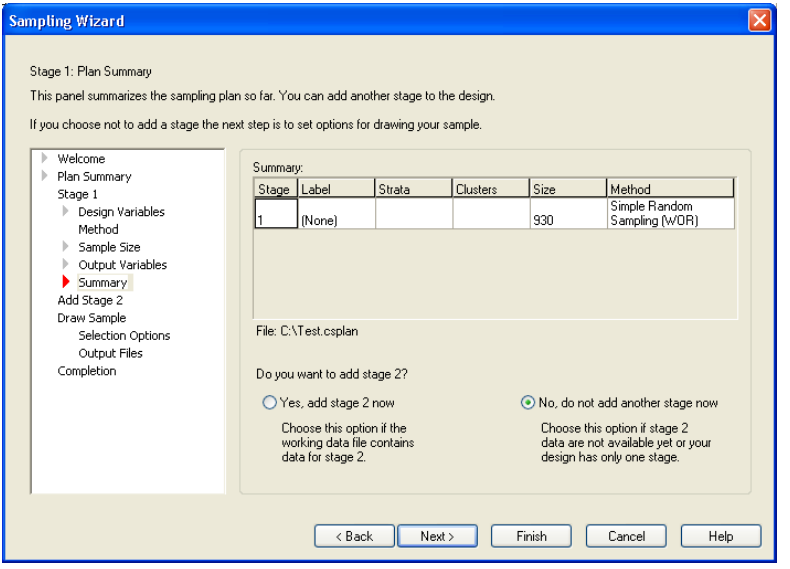


Figura 8.10 – Pasul 4: Sumarul stadiului

Noua etapă afişează sumarul stadiului curent. Obținem scurte informații despre numărul stadiului, eticheta acestuia – în cazul în care am definit una – variabila de stratificare, variabila de cluster, dimensiunea eşantionului şi metoda de eşantionare folosită. La acest nivel avem posibilitatea creării unui nou stadiu selectând opțiunea **„Yes, add stage 2 now”** în cazul în care

ne interesează o eşantionare multistadială sau iniţierea etapelor de creare efectivă a eşantionului alegând **„No, do not add another stage now”** care este de fapt şi varianta implicit selectată. De asemenea, ne putem întoarce la etapele anterioare apăsând butonul **„Back”**.

Pentru moment, nu suntem interesaţi decât de crearea unui eşantion simplu aleatoriu, fără alte stadii suplimentare, astfel încât vom apăsa butonul **„Next”**.

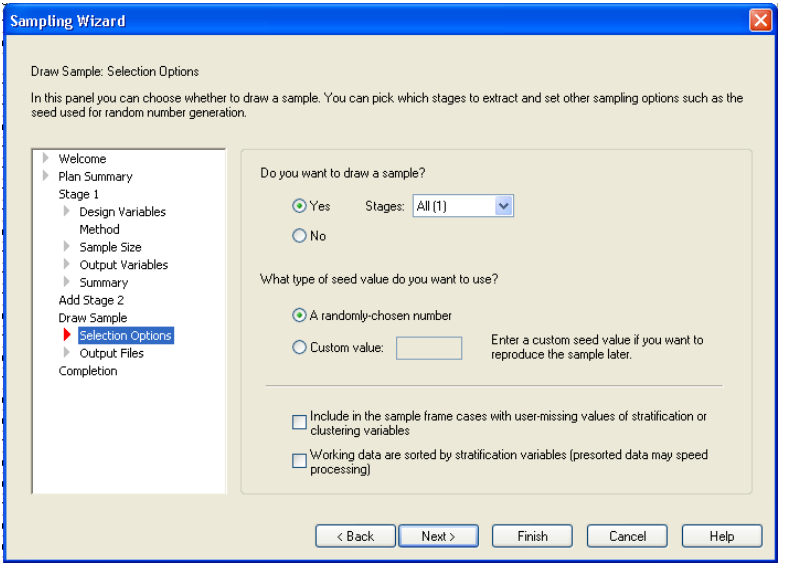

Sampling Wizard
Draw Sample: Selection Options
In this panel you can choose whether to draw a sample. You can pick which stages to extract and set other sampling options such as the seed used for random number generation.
Welcome
Plan Summary
Stage 1
Design Variables
Method
Sample Size
Output Variables
Summary
Add Stage 2
Draw Sample
Selection Options
Output Files
Completion
Do you want to draw a sample?
Yes
Stages:
All (1)
No
What type of seed value do you want to use?
A randomly-chosen number
Custom value:
Enter a custom seed value if you want to reproduce the sample later.
Include in the sample frame cases with user-missing values of stratification or clustering variables
Working data are sorted by stratification variables (presorted data may speed processing)
< Back
Next >
Finish
Cancel
Help

Figura 8.11 – Pasul 5: Crearea eşantionului. Opţiuni de selecţie

Noua fereastră prezintă o serie de opţiuni ce iniţiază modalitatea de creare efectivă a eşantionului. Sistemul ne întreabă dacă dorim să creăm acum eşantionul (**„Do you want to draw a sample?”**). Putem alege crearea eşantionului, selectând **„Yes”**, situaţie în care vom avea posibilitatea de generare a tuturor stadiilor, alegând **„All”** din lista derulantă sau crearea eşantionului doar pentru un anumit stadiu, prin selectarea numărului acestuia. Stadii-

le pot fi generate doar în ordine. Astfel, nu putem crea stadiul 3 în condițiile în care nu au fost executate stadiile 1 şi 2.

Secțiunea **„What type of seed value do you want to use?"** permite stabilirea valorii de inițializare pentru generatorul de numere aleatoare. Putem alege între un număr oarecare, generat automat de computer (selectând **„A randomly-chosen number"**) sau putem include o valoare în caseta de text **„Custom value"** în situația în care se doreşte reproducerea caracteristicilor eşantionului.

Cele două casete de bifare permit tratarea cazurilor în care lipsesc informații la nivelul unor elemente şi accelerarea procesului de generare a eşantionului. Bifarea primei casete are ca efect includerea într-o categorie separată a unităților cu date lipsă iar debifarea acesteia ignoră unitățile care au date lipsă.

Bifarea celei de-a doua casete are ca efect accelerarea procesului de generare a eşantionului în condițiile în care datele sunt deja sortate după valorile unei variabile de stratificare. Altminteri, SPSS va proceda la o nouă resortare, operațiune consumatoare de timp.

Următoarea etapă permite precizarea locului în care va fi salvat eşantionul şi variabilele pe care acesta le generează. Putem opta între baza de date curentă (**„Active dataset"**), caz în care SPSS va selecta din baza de eşantionare unitățile incluse în eşantion şi va calcula variabilele necesare doar pentru acestea, putem alege o nouă bază de date (**„New dataset"**), caz în care SPSS va crea o nouă bază de date, nesalvată, în care va include doar unitățile selectate în eşantion sau putem crea o nouă bază de date cu salvare (**„External file"**), caz similar celui anterior, singura deosebire fiind aceea că SPSS va crea un nou fişier pe disc şi va salva eşantionul. În cele două situații va trebui să precizăm numele bazei de date, respectiv numele şi calea către fişierul de date.

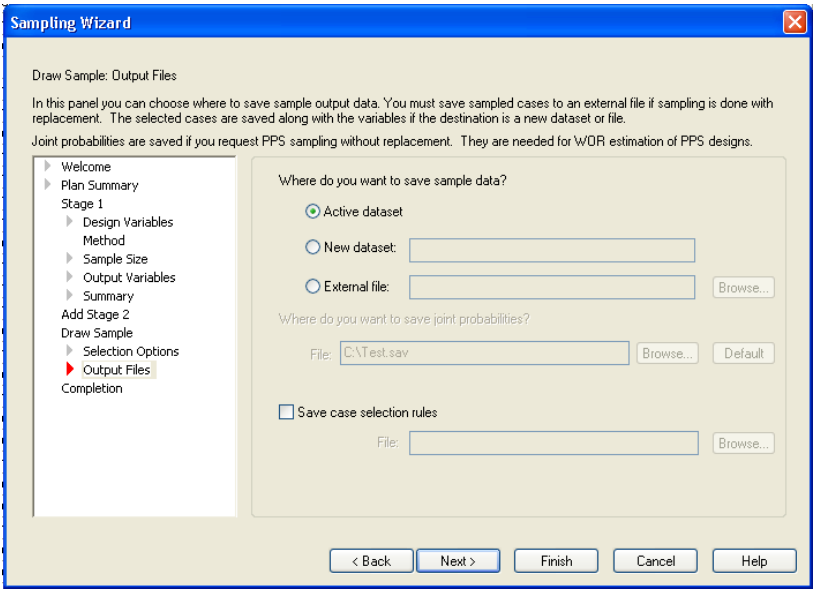


Figura 8.12 – Pasul 5: Crearea eşantionului. Opţiuni de salvare

În această fereastră se mai află un număr de două elemente:

- Fişierul în care vor fi salvate probabilităţile cumulate (**„Joint probabilities”**) în cazul în care se folosesc eşantioane proporţionale cu mărimea şi fără introducerea bilei în urnă
- Regulile de selecţie ale cazurilor (**„Case selection rules”**) în cazul în care se generează un singur stadiu la un moment dat şi se doreşte salvarea regulilor de selecţie a cazurilor într-un fişier separat. Opţiunea este utilă la construirea sub-stadiilor.

În situaţia noastră, am ales salvarea eşantionului în aceeaşi bază de date pentru a putea urmări mai uşor procesul de selecţie. Apăsarea butonului **„Next”** permite trecerea la ultima etapă a creării eşantionului.

Ultima fereastră ne întreabă dacă dorim să salvăm designul realizat în planul de eşantionare şi să creăm efectiv eşantionul (opțiunea **„Save the design to a plan and draw the sample”**) sau dorim să copiem într-un fişier de sintaxă comenzile SPSS generate în urma prelucrărilor noastre (opțiunea **„Paste de syntax generated by the Wizard into a syntax window”**).

Evident, vom alege prima opțiune şi vom apăsa apoi butonul **„Finish”.**

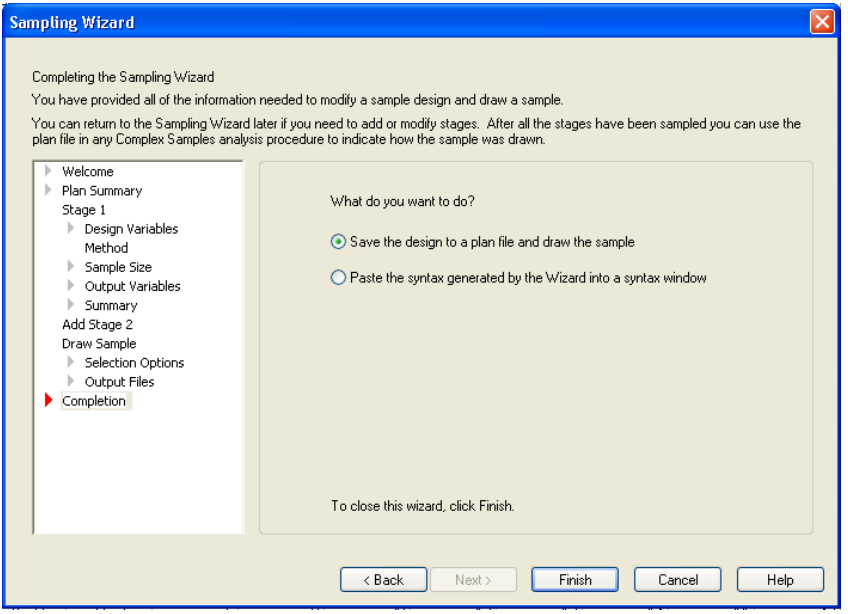

Sampling Wizard
Completing the Sampling Wizard
You have provided all of the information needed to modify a sample design and draw a sample.
You can return to the Sampling Wizard later if you need to add or modify stages. After all the stages have been sampled you can use the plan file in any Complex Samples analysis procedure to indicate how the sample was drawn.
Welcome
Plan Summary
Stage 1
Design Variables
Method
Sample Size
Output Variables
Summary
Add Stage 2
Draw Sample
Selection Options
Output Files
Completion
What do you want to do?
Save the design to a plan file and draw the sample
Paste the syntax generated by the Wizard into a syntax window
To close this wizard, click Finish.
< Back
Next >
Finish
Cancel
Help

Figura 8.13 – Pasul 6: Crearea eşantionului

După câteva fracțiuni de secundă, SPSS procedează la crearea eşantionului şi ne oferă o serie de informații în fereastra de rezultate (Output).

Primul tabel din fereastra de rezultate se referă la informații despre planul de eşantionare. Observăm că avem un singur stadiu, metoda de selecție este simplu aleatorie fără reintroducerea bilei în urnă iar eşantionul conți-

ne un număr de 930 de cazuri. În acest tabel regăsim informaţii despre noile variabile create în baza de date şi, de asemenea, ne este furnizat fişierul ce conţine planul de eşantionare.

Tabelul planului de eşantionare prezintă şi o serie de date utile în proiectarea planului de analiză a eşantionului pe care-l vom discuta în următorul subcapitol.

O variabilă foarte importantă la acest nivel este variabila *„SampleWeight_Final_”,* folosită în general la proiectarea planului de analiză a eşantionului ca variabilă caracteristică a reprezentativităţii cazurilor în eşantion.

Summary

			Stage 1
Sample Information	Selection Method		Simple random sampling without replacement
	Number of Units Sampled		930
	Variables Created or Modified	Stagewise Inclusion (Selection) Probability	InclusionProbability_1_
		Stagewise Cumulative Sample Weight	SampleWeight Cumulative_1_
		Stagewise Population Size	PopulationSize_1_
		Stagewise Sample Size	SampleSize_1_
		Stagewise Sampling Rate	SamplingRate_1_
		Stagewise Sample Weight	SampleWeight_1_
Analysis Information	Estimator Assumption		Equal probability sampling without replacement
	Inclusion Probability		Obtained from variable InclusionProbability_1_

Plan File: C:\Test.csplan
Weight Variable: SampleWeight_Final_

Figura 8.14 – Planul de eşantionare

Summary for Stage 1

Number of Units Sampled		Proportion of Units Sampled	
Requested	Actual	Requested	Actual
930	930	61,3%	61,3%

Plan File: C:\Test.csplan

Figura 8.15 – Sumarul stadiului

Următorul tabel generat se referă la sumarul stadiului. Pentru fiecare sta-

diu al eşantionării va fi generat un asemenea tabel, în care regăsim informaţii despre unităţile şi proporţiile selectate.

Observăm din nou numele şi calea fişierului de tip plan de eşantionare. În cazul nostru am solicitat includerea unui număr de 930 de cazuri în eşantion (**Requested**) iar SPSS a reuşit selectarea tuturor subiecţilor (**Actual**). Procentul solicitat a fost de 61,3%, procent atins în final de către SPSS.

517 — Visible: 50 of 50 Variables

4	InclusionProbability_1_	SampleWeightCumulative_1_	PopulationSize_1	SampleSize_1_	SamplingRate_1_	SampleWeight_1_	SampleWeight_Final_
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	,61	1,63	1517	930	,61	1,63	1,63
.	.	.	.	.	.	.	.
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63
.	,61	1,63	1517	930	,61	1,63	1,63

Figura 8.16 – Noile variabile incluse în baza de date

Trecând la baza de date, observăm la sfârşitul acesteia variabilele nou create. Vom constata şi selecţia eşantionului făcută de SPSS. Programul a ales în eşantion primii şapte subiecţi, pe următorii şase nu i-a selectat, a selectat apoi un subiect, următorul nefiind selectat şi aşa mai departe.

Luând fiecare subiect în parte, observăm o probabilitate de includere egală de 0,61, dintr-o populaţie de 1517 la mărimea unui eşantion de 930.

Variabila *„SampleWeight”* este generată automat şi corespunde oarecum frecvenţei cu care fiecare caz selectat în eşantion este reprezentat în populaţia generală. Suma acestor valori se apropie foarte mult de dimensiunea populaţiei. Într-adevăr, avem 930 de cazuri, fiecare având valoarea 1,63 la

această variabilă. Prin urmare, 930 × 1,63 = 1515,9 valoare apropiată de mărimea populației. Această variabilă este foarte importantă în procesul de analiză a eşantionului.

Iată că, urmând aceşti paşi, ați reuşit crearea unui eşantion. Realizarea acestuia nu oferă însă informații legate de caracteristicile pe care le poate avea. După selectarea subiecților în eşantion urmează etapa de analiză a eşantionului în vederea identificării proprietăților sale statistice.

VIII.4.2 Analiza eşantionului

Crearea unui plan de analiză reprezintă prima etapă în studiul proprietăților unui eşantion. Vom folosi din nou meniul **„Analyze”**, apoi submeniul **„Complex Samples”** şi în final vom efectua click pe opțiunea **„Prepare for Analysis…”**

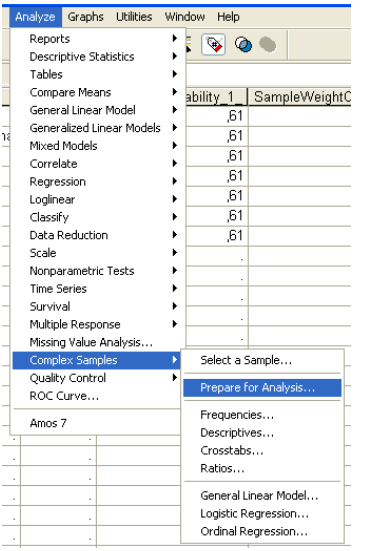


Figura 8.17 – Meniul planului de analiză

Accesarea acestei comenzi determină lansarea în execuție a unui nou asistent (Wizard), specializat în proiectarea planului de analiză, al doilea plan ca importanță în studiul eşantionului.

Planul de analiză, similar planului de eşantionare, reprezintă o succesiune de paşi prin intermediul cărora se creează premisele studiului proprietăților unui eşantion. După crearea unui eşantion este importantă analiza acestuia în vederea determinării caracteristicilor sale.

Primul pas este reprezentat de specificarea fişierului în care va fi salvat planul de analiză. Opțiunea **„Create a plan file”** ne permite alegerea locului şi stabilirea numelui sub care va fi salvat planul de analiză. Puteți tasta

calea completă sau puteți folosi butonul **„Browse…”** prin intermediul căruia veți alege locația acestui fișier. Evident, va trebui să furnizați și un nume. În cazul nostru, fișierul plan de analiză se numește *„Analiza.csaplan”* și a fost salvat direct în directorul *C:*

Observați diferența de extensie. Dacă planurile de eșantionare aveau extensia *.csplan*, planul de analiză are extensia *.csaplan*.

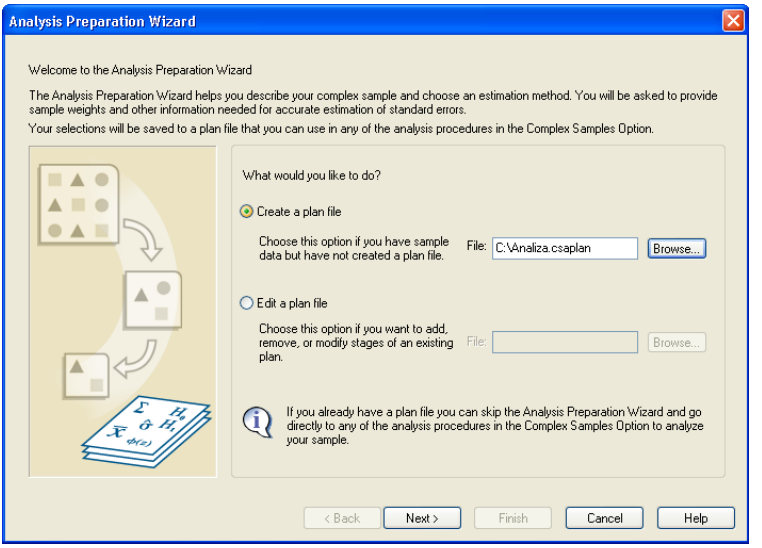

Analysis Preparation Wizard
Welcome to the Analysis Preparation Wizard
The Analysis Preparation Wizard helps you describe your complex sample and choose an estimation method. You will be asked to provide sample weights and other information needed for accurate estimation of standard errors.
Your selections will be saved to a plan file that you can use in any of the analysis procedures in the Complex Samples Option.
What would you like to do?
Create a plan file
Choose this option if you have sample data but have not created a plan file.
File: C:\Analiza.csaplan
Browse...
Edit a plan file
Choose this option if you want to add, remove, or modify stages of an existing plan.
File:
Browse...
If you already have a plan file you can skip the Analysis Preparation Wizard and go directly to any of the analysis procedures in the Complex Samples Option to analyze your sample.
< Back
Next >
Finish
Cancel
Help

Figura 8.18 – Pasul 1. Salvarea planului de analiză

Opțiunea **„Edit a plan file”** vă permite modificarea configurației unui plan de analiză salvat anterior. Apăsarea butonului **„Next”** determină trecerea la următoarea etapă.

Noua fereastră conține mai multe elemente. În primul rând va trebui să specificăm variabila caracteristică „reprezentativității” eșantionului. Am accentuat deja asupra importanței variabilei *„Sample Weight”* tocmai în ideea de a fi familiarizați cu aceasta. Deși în cazul unor eșantioane complexe

această variabilă poate lua diferite forme, în planul de analiză vom include, de obicei, forma finală a acesteia, forma în care SPSS ponderează de fapt toate stadiile. Vă recomand introducerea în secțiunea **„Sample Weight”** a variabilei de tipul *„Final Sample Weight”*, variabilă care de obicei caracterizează cel mai bine eşantionul. În cazul planurilor complexe, s-ar putea include şi alte variabile parțiale de acest tip dacă se doreşte investigarea la nivelul stadiilor, a straturilor ori a clusterilor.

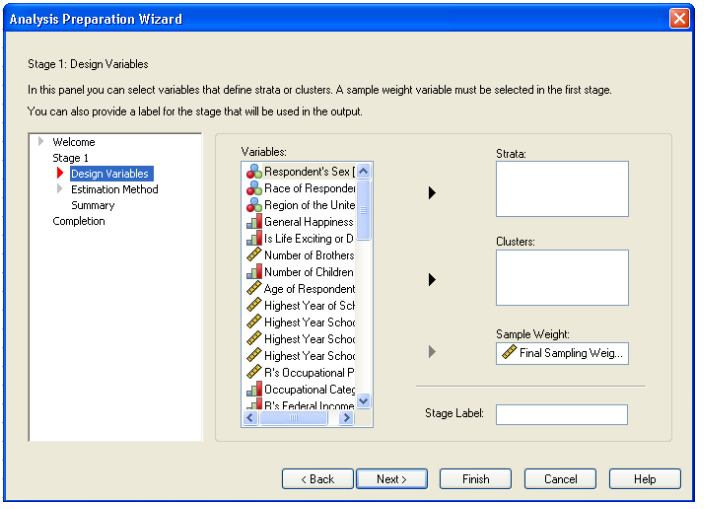


Figura 8.19 – Pasul 2. Stabilirea variabilelor

Alte posibilități cuprinse la acest nivel pot fi reprezentate de:

- Variabilele de stratificare (**„Strata”**) – va fi inclusă variabila de stratificare în condițiile unui eşantion stratificat. Eşantionul total va fi reprezentat din ponderile pe care le au cazurile particulare din fiecare strat;
- Variabilele cluster (**„Clusters”**) – se include variabila ce defineşte clusterii, grupuri de unități de observație în cazul în care se analizează un eşantion pe clusteri;

- **„Stage Label”** – permite definirea unei etichete a stadiului de analiză, similar procesului de creare a planului de eşantionare.

Apăsarea butonului **„Next”** determină trecerea la alegerea metodei de estimare.

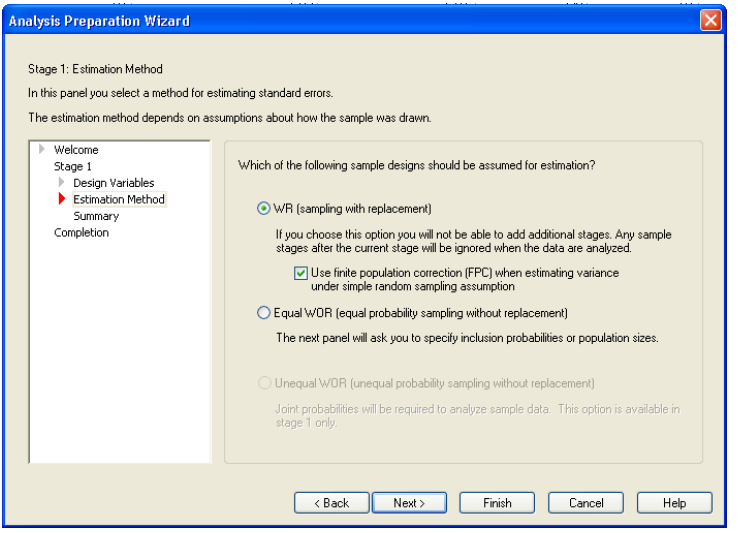

Analysis Preparation Wizard
Stage 1: Estimation Method
In this panel you select a method for estimating standard errors.
The estimation method depends on assumptions about how the sample was drawn.
Welcome
Stage 1
Design Variables
Estimation Method
Summary
Completion
Which of the following sample designs should be assumed for estimation?
WR (sampling with replacement)
If you choose this option you will not be able to add additional stages. Any sample stages after the current stage will be ignored when the data are analyzed.
Use finite population correction (FPC) when estimating variance under simple random sampling assumption
Equal WOR (equal probability sampling without replacement)
The next panel will ask you to specify inclusion probabilities or population sizes.
Unequal WOR (unequal probability sampling without replacement)
Joint probabilities will be required to analyze sample data. This option is available in stage 1 only.
< Back
Next >
Finish
Cancel
Help

Figura 8.20 – Pasul 2. Alegerea metodei de estimare

Alegerea metodei de estimare permite specificarea doar a tipului de selecție utilizat. În funcție de modul de proiectare a eşantionului puteți opta pentru metoda „cu introducerea bilei în urnă” (**WR**) în care nu vor fi incluse corecții pentru eşantioane provenite din populații finite la estimarea varianței. Forțarea includerii acestor corecții se realizează prin bifarea casetei **FPC** care se activează la alegerea acestei opțiuni.

Metoda „cu includerea bilei în urnă” (**WOR**) poate fi selectată în cazul eşantioanelor cu probabilitate egală de selecție şi fără înlocuire. Această metodă include corecțiile pentru populații finite şi pleacă de la supoziția conform căreia unitățile au fost selectate cu o probabilitate egală.

Metoda „cu includerea bilei în urnă” (**WOR**) şi cu probabilitate inegală foloseşte şi ea corecţiile de populaţie finită dar pleacă de la supoziţia că selecţia unităţilor din populaţie s-a realizat cu o probabilitate inegală. În general metoda se foloseşte în cazul unor eşantioane nealeatorii.

Eşantionul nostru a fost creat fără „includerea bilei în urnă” şi cu o probabilitate egală, prin urmare vom alege opţiunea a doua.

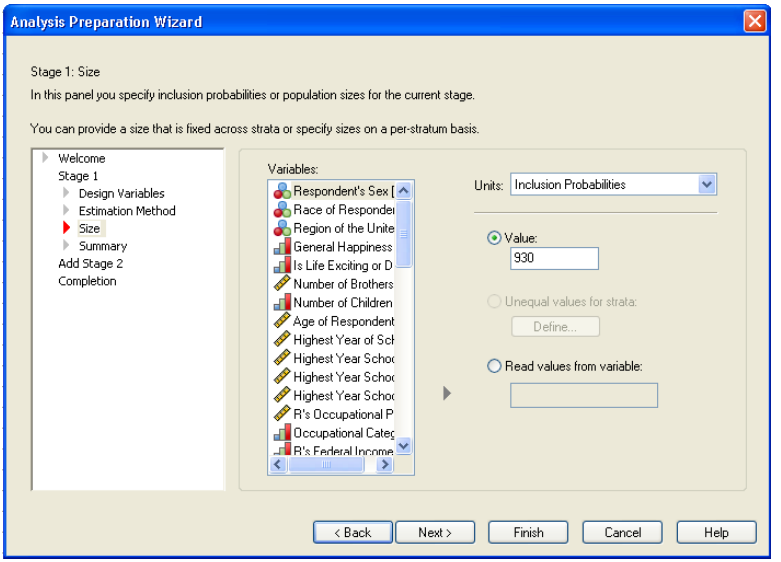


Figura 8.21 – Pasul 2. Alegerea dimensiunii

Apăsarea butonului **„Next”** vă permite să specificaţi probabilităţile de includere sau mărimea populaţiei la nivelul stadiului curent. Aceste dimensiuni pot fi fixe sau pot varia la nivelul straturilor. Această etapă este valabilă doar dacă anterior aţi ales opţiunea WOR cu probabilităţi egale.

Folosind caseta de derulare **„Units”** puteţi preciza dimensiunea exactă a populaţiei sau probabilitatea cu care elementele au fost extrase. În primul caz putem introduce în caseta de text **„Value”** cifra 1517, valoare ce repre-

zintă dimensiunea exactă a populației sau 0,61, probabilitatea de includere. Am preferat, în exemplul nostru, prima variantă. Utilizarea acestei secțiuni presupune o singură valoare aplicată tuturor straturilor.

O altă variantă o reprezintă opțiunea **„Unequal values for strata”** care se activează numai în situația definirii anterioare a unei variabile de stratificare. Caseta ce se deschide la apăsarea butonului **„Define…”** permite stabilirea valorilor pentru fiecare strat.

Ultima opțiune, **„Read values from variable”** ne permite includerea unei variabile numerice care conține valorile pentru fiecare strat.

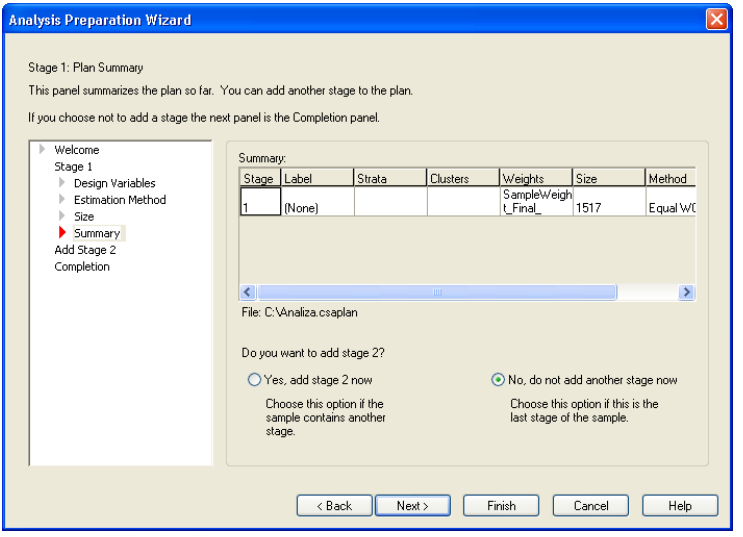


Figura 8.22 – Pasul 2. Sumarul planului

Similar planului de eşantionare, următoarea fereastră prezintă un sumar al planului de analiză şi ne permite construirea unor noi stadii. Nu vom intra în detalii referitoare la această secțiune deoarece am tratat complet problema în cadrul capitolului referitor la planul de eşantionare.

Ultimul ecran ne invită să salvăm planul de analiză creat sau să copiem sintaxa într-o fereastră de comenzi SPSS, într-un mod identic cu planul de eşantionare.

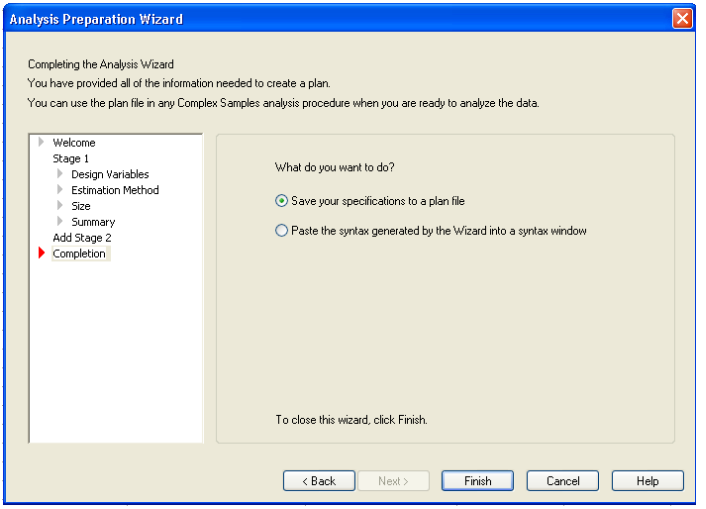

Analysis Preparation Wizard
Completing the Analysis Wizard
You have provided all of the information needed to create a plan.
You can use the plan file in any Complex Samples analysis procedure when you are ready to analyze the data.
Welcome
Stage 1
Design Variables
Estimation Method
Size
Summary
Add Stage 2
Completion
What do you want to do?
Save your specifications to a plan file
Paste the syntax generated by the Wizard into a syntax window
To close this wizard, click Finish.
< Back
Next >
Finish
Cancel
Help

Figura 8.23 – Pasul 3. Salvarea planului de analiză

Un plan de analiză salvat va putea fi utilizat în vederea investigării proprietăţilor eşantionului. La apăsarea butonului **„Finish”**, planul de analiză va fi salvat iar în fereastra de rezultate va fi redată structura acestuia.

VIII.4.3 Investigarea proprietăţilor eşantionului

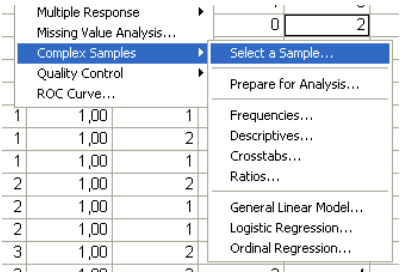


Figura 8.24 – Analiza eşantionului

Odată stabilite planul de eşantionare şi planul de analiză se poate trece la etapa finală a eşantionării şi anume investigarea proprietăţilor eşantionului.

SPSS pune la dispoziţie mai multe metode de investigare a unui eşantion. Deoarece nu dispunem încă de toate informaţiile necesa-

re studiului exhaustiv al acestor proceduri, vom limita expunerea doar la două proceduri ce implică statistici univariate, după cum urmează:

- Meniul **„Frequencies…”** determină constituirea tabelelor de frecvențe pentru variabilele selectate şi afişarea statisticilor de tip univariat. Procedura estimează frecvențele eşantionului şi erorile standard, intervalele de încredere, coeficienții de variație şi ale statistici utile studiului unui eşantion;
- Meniul **„Descriptives…”** produce de asemenea statistici univariate de tipul celor menționate existând şi posibilitatea analizei după subgrupuri;

Accesarea primului meniu, **„Frequencies…”** determină afişarea primului ecran necesar analizei prin care urmează să alegem planul de eşantionare dorit.

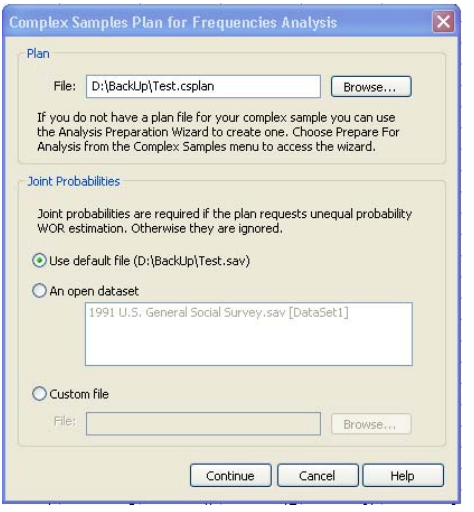

Complex Samples Plan for Frequencies Analysis
Plan
File:
D:\BackUp\Test.csplan
Browse...
If you do not have a plan file for your complex sample you can use the Analysis Preparation Wizard to create one. Choose Prepare For Analysis from the Complex Samples menu to access the wizard.
Joint Probabilities
Joint probabilities are required if the plan requests unequal probability WOR estimation. Otherwise they are ignored.
Use default file (D:\BackUp\Test.sav)
An open dataset
1991 U.S. General Social Survey.sav [DataSet1]
Custom file
File:
Browse...
Continue
Cancel
Help

Figura 8.25 – Alegerea planului de eşantionare

Toate operaţiile din această etapă impun anumite specificaţii de analiză, specificaţii conţinute în planul de eşantionare, după cum am văzut în subcapitolul anterior. În secţiunea **„Plan”**, caseta de text **„File”** vom putea tasta calea către planul de eşantionare sau vom putea utiliza butonul **„Browse…”** în vederea navigării către locaţia ce conţine acest fişier.

În cazul în care probabilităţile de includere ale elementelor în eşantion nu sunt egale, situaţie des întâlnită în cadrul eşantionării stratificate sau a eşantionării pe clusteri, va trebui să specificăm fişierul în care s-au precizat aceste probabilităţi, utilizând secţiunea **„Joint Probabilitoes**”. Astfel, putem alege aceste probabilităţi dintr-o variabilă din baza de date, dintr-o altă bază de date sau dintr-un alt tip de fişier.

Trecerea la analiza efectivă se face prin apăsarea butonului **„Continue”**.

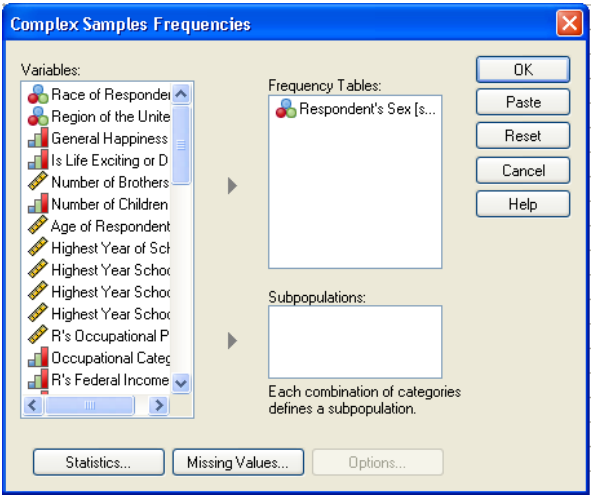


Figura 8.26 – Stabilirea variabilelor de interes

Va apărea, iată, o fereastră pe care o cunoaşteţi destul de bine. Acest ecran permite stabilirea variabilei pe care dorim să o analizăm. Se impune o

mențiune la acest nivel: variabilele folosite pot fi doar variabile categoriale, situate la un nivel de măsură nominal sau ordinal. În partea stângă a ferestrei se află secțiunea **„Variables”**, o listă ce conține toate variabilele din baza de date. Lista **„Frequency Tables”** permite includerea variabilelor pentru care dorim să calculăm tabele de frecvență. Cu ale cuvinte, variabilele pe care dorim să le analizăm. În exemplul nostru, vom dori analiza eşantionului sub aspectul sexului (al genului biologic). Vom transfera aşadar această variabilă în lista **„Frequency Tables”** prin utilizarea butonului de transfer.

Lista **Subpopulations”** permite includerea unei alte variabile categoriale în condițiile în care am realizat, spre exemplu, un eşantion stratificat. Nu este cazul nostru, însă dacă am fi dorit să analizăm reprezentativitatea genului biologic în funcție de mediul de proveniență (rural vs. urban) am fi inclus această ultimă variabilă în lista **„Subpopulations”**. În această situație, SPSS ar fi analizat separat bărbații şi femeile în funcție de mediul lor de proveniență.

Un buton deosebit de important este butonul **„Statistics...”**. Apăsarea acestuia determină posibilitatea stabilirii indicatorilor statistici ce vor fi calculați.

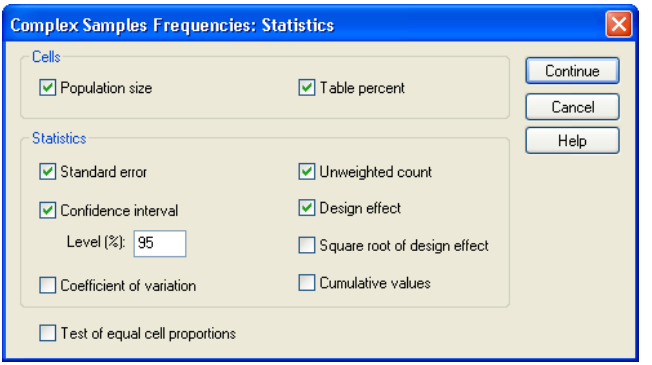


Figura 8.27 – Stabilirea indicatorilor statistici

Există destul de multe opțiuni în cadrul acestei noi ferestre, unele dintre ele fiindu-vă necunoscute deoarece nu aveți încă informațiile necesare referitoare la anumite teste statistice. Noi le vom prezenta pe toate, studiind aici în detaliu doar câteva, celelalte urmând să le analizăm în volumele ulterioare.

Secțiunea **„Cells”** permite afişarea informațiilor legate de populația țintă din care a fost extras eşantionul. Vor fi afişate date referitoare la dimensiunea populației sub formă de frecvențe absolute prin bifarea casetei **„Population size”** şi sub formă de frecvențe relative (procente) bifând caseta **„Table percent”**.

Secțiunea **„Statistics”** permite configurarea indicatorilor statistici asociați dimensiunii populației sau a procentului din populație, astfel:

- Prin bifarea casetei **„Standard error”** vom calcula eroarea standard a estimării;
- Caseta **„Confidence interval”** permite stabilirea intervalului de încredere a estimării în baza nivelului specificat. În mod normal, intervalul de încredere îl vom stabili la 95%, după cum observați că am procedat şi în exemplul nostru (caseta de text **„Level%”**);
- Caseta **„Coefficient of variation”** comunică programului calculul coeficientului de variație a estimării, sub formă de proporții;
- Caseta **„Unweighted count”** permite afişarea numărului de elemente folosit în realizarea estimării. Cu alte cuvinte, dimensiunea eşantionului extras;
- Caseta **„Design effect”** permite calculul variației estimării. Indicatorul exprimă proporția de varianță a eşantionului comparată cu varianța unui eşantion simplu randomizat. Coeficien-

tul nu este folosit în cazul eşantioanelor simplu randomizate – în acest caz fiind întotdeauna 1 – ci doar în cazul eşantioanelor mai complexe. Cu cât valoarea este mai depărtată de unu, cu atât efectul este mai mare şi deci varianța este mai mare;

- Caseta **„Square root of design effect"** reprezintă rădăcina pătrată din varianța estimării (un fel de abatere standard dacă vă mai aduceți aminte), fiind o măsură standard a efectului şi având acelaşi mod de interpretare;
- Caseta **„Cumulative values"** calculează frecvența cumulată a populației țintă, având aceeaşi semnificație ca orice frecvență cumulată.
- Caseta **„Test of equal cell proportions"** permite calculul a două teste statistice pe care nu le-am studiat: testul χ^2 şi testul de probabilitate a proporțiilor. Fără a intra în amănunte, menționăm că ambele teste pornesc de la ipoteza nulă că toate categoriile unei variabile au frecvențe egale. Testele ne permit să verificăm dacă există diferențe semnificative între frecvențele categoriilor variabilei (în cazul nostru dacă frecvența bărbaților este semnificativ mai mare sau semnificativ mai mică în comparație cu frecvența femeilor)

Pentru exercițiul nostru am comunicat programului SPSS să afişeze frecvența absolută şi relativă a populației țintă (casetele **„Population size"** şi **„Table percent"**), să calculeze eroarea standard a estimării, intervalul de încredere şi – doar demonstrativ – mărimea efectului şi să afişeze mărimea eşantionului (casetele **„Standard error"**, **„Confidence interval"** la un nivel de 95%, **„Unweighted count"** şi **„Design effect"**). Apăsând butonul **„Continue"** revenim la fereastra inițială.

Un alt buton prezent este butonul **„Missing Values”** prin apăsarea căruia vom putea stabili modul de tratare a cazurilor lipsă.

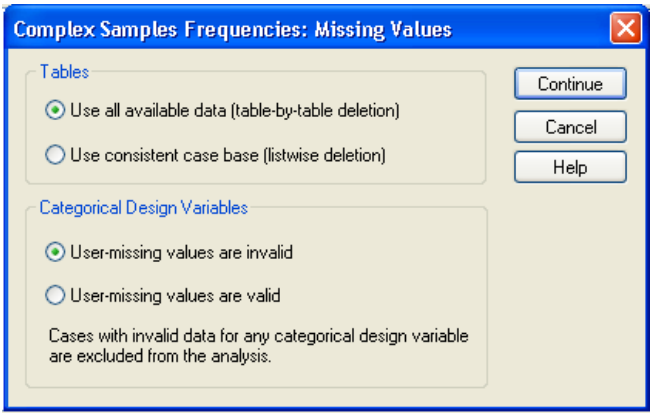

Complex Samples Frequencies: Missing Values
Tables
Use all available data (table-by-table deletion)
Use consistent case base (listwise deletion)
Continue
Cancel
Help
Categorical Design Variables
User-missing values are invalid
User-missing values are valid
Cases with invalid data for any categorical design variable are excluded from the analysis.

Figura 8.28 – Tratarea cazurilor lipsă

În secţiunea **„Tables”** putem stabili ce cazuri vor fi incluse în analiză. Opţiunea implicită este **„Use all variable data”** în care vor fi analizate doar valorile valide la nivelul fiecărei variabile. Spre exemplu, dacă analizăm variabila „gen biologic”, SPSS va urmări să existe valori valide doar la nivelul acestei variabile, fără a fi preocupat dacă există valori valide la alte variabile.

Opţiunea **„Use consistent case base”** este însă mult mai restrictivă. Va trebui să existe valori la nivelul tuturor variabilelor analizate pentru ca subiectul respectiv să fie luat în considerare. De exemplu, dacă analizăm variabilele „gen biologic” şi „mediu de proveniență”, un subiect va trebui să aibă scoruri valide la ambele variabile, chiar dacă este analizată doar variabila „gen biologic”. Un subiect care are un scor valid la „gen biologic” dar nu are date la „mediul de proveniență” va fi exclus din analiză.

Secţiunea **„Categorical design variables”** determină, în cazul variabilelor categoriale, dacă valorile lipsă vor fi tratate ca fiind sau nefiind valide.

Practic nu este nevoie să intervenim în această fereastră decât extrem de rar şi în cazul unor planuri de eşantionare complexe. Apăsarea butonului **„Continue”** va determina şi în acest caz întoarcerea la fereastra inițială.

Lansarea efectivă a analizei datelor se face prin apăsarea butonului **„OK”.** În câteva fracțiuni de secundă va fi afişat un tabel în fereastra de rezultate, tabel ce conține exact datele solicitate.

Respondent's Sex

		Estimate	Standard Error	95% Confidence Interval		Design Effect	Unweighted Count
				Lower	Upper		
Population Size	Male	662,260	15,355	632,126	692,395	1,000	406
	Female	854,740	15,355	824,605	884,874	1,000	524
	Total	1517,000	,000	1517,000	1517,000	.	930
% of Total	Male	43,7%	1,0%	41,7%	45,7%	1,000	406
	Female	56,3%	1,0%	54,3%	58,3%	1,000	524
	Total	100,0%	,0%	100,0%	100,0%	.	930

Figura 8.29 – Rezultatele analizei eşantionului

Capul de tabel conține numele variabilei de interes şi denumirea indicatorilor calculați. Primul rând se referă la frecvențele absolute ale populației țintă iar al doilea rând la frecvențele relative. Observăm că datele noastre au fost împărțite pe cele două categorii ale variabilei de interes – bărbați şi femei.

A doua coloană (**Estimate**) exprimă frecvențele absolute şi relative ale populației țintă estimată în baza eşantionului. Aceste elemente nu reprezintă date reale ci date estimate din eşantion. Nu putem avea un număr de 662,26 bărbați şi 854,74 femei. Un bărbat sau o femeie este un întreg, nu are zecimale. Observăm însă că dimensiunea totală a populației țintă estimate (1517 subiecți) este egală cu populația țintă reală. Exprimat procentual, eşantionul nostru estimează o populație compusă din 43,7% bărbați şi 56,3% femei.

Următoarea coloană (**Standard Error**) se referă la eroarea standard a estimării, exprimată atât în formă brută cât şi în formă procentuală. Eşantionul nostru estimează populaţia ţintă cu o eroare standard de 15,35 bărbaţi şi 15,35 femei (aproximativ 15 subiecţi pentru fiecare categorie, cifrele reprezentând din nou o abstracţiune). Acest indicator devine mai clar dacă urmărim forma procentuală. Iată că eroarea estimării este de 1% atât în cazul bărbaţilor cât şi în cazul femeilor, care la un nivel de încredere de 95% ne poate determina să afirmăm că avem un eşantion reprezentativ sub aspectul sexului.

Urmează două coloane ale intervalului de încredere. În baza erorii standard, eşantionul estimează practic o populaţie ţintă formată dintr-un număr de 632,12 până la 692,39 bărbaţi şi de la 824,60 până la 884,87 femei. Cu alte cuvinte, compoziţia populaţiei ţintă estimată de eşantion este formată din 41,7 până la 45,7 bărbaţi şi 54,3 până la 58,3% femei. Undeva între aceste limite putem găsi populaţia ţintă reală estimată de eşantion.

Coloana **Design Effect** are evident valoarea 1 deoarece am lucrat cu un eşantion simplu randomizat, fapt tratat mai sus.

Ultima coloană indică dimensiunea şi compoziţia eşantionului (**Unweighted Count**). După cum ştiam deja, eşantionul are un număr de 930 de cazuri iar din punct de vedere al sexului avem 406 bărbaţi şi 524 femei. Iată că acest eşantion estimează reprezentativ populaţia ţintă la o eroare de doar 1% şi la un nivel de încredere de 95%. Suntem în faţa unui eşantion reprezentativ, acest lucru fiind demonstrat prin analiza efectuată.

Am analizat o variabilă de interes situată la nivel nominal (sexul). Cum procedăm însă dacă avem o variabilă de interes la un nivel scalar, de exemplu vârsta? Analiza frecvenţelor nu ne ajută prea mult. Este timpul să abordăm cea de-a doua metodă propusă şi anume analiza descriptivă. Paşii sunt identici, astfel încât nu vom insista asupra lor. Există însă o serie de metode statistice uşor diferite pe care vom încerca să le lămurim.

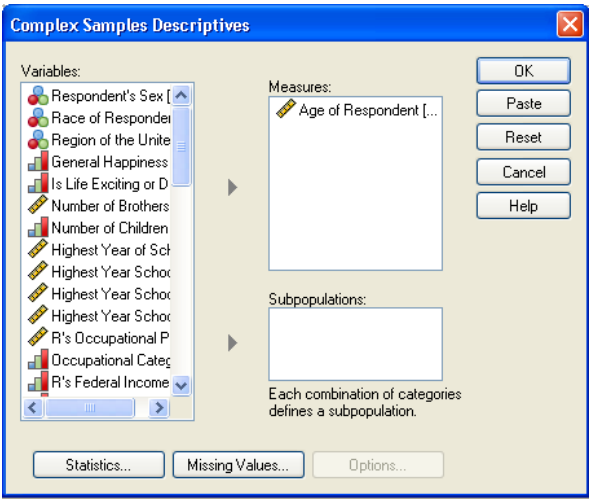


Figura 8.30 – Analiza variabilei de interes de tip scalar

După lansarea procedeului de analiză şi alegerea planului de eşantionare prin apelarea meniului **„Descriptives”**, va trebui să includem, similar procesului anterior, variabila de interes. În cazul nostru am ales o variabilă scalară, vârsta subiecților. Apăsarea butonului **„Statistics”** determină apariția unui ecran uşor diferit în comparație cu analiza frecvențelor

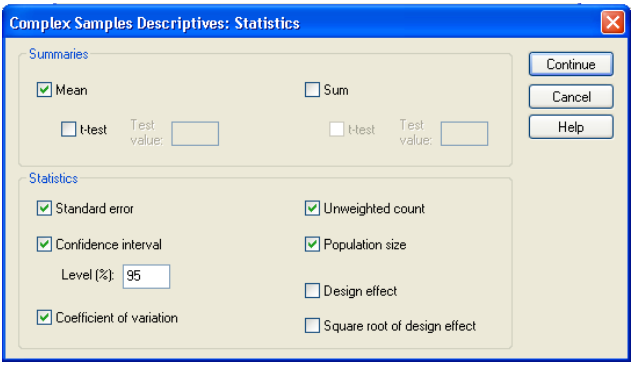


Figura 8.31 – Stabilirea indicatorilor statistici

Diferența rezidă în secțiunea **„Summaries”** prin intermediul căreia putem estima mediile şi sumele variabilelor țintă. În cazul în care media populației este cunoscută (uneori se foloseşte media teoretică) putem folosi testele „t” Student de comparație între media populației şi media unui eşantion pentru verificarea existenței unei diferențe semnificative între medii. Despre aceste teste vom discuta însă în volumul următor.

Celelalte opțiuni ale acestei ferestre au fost discutate şi nu este cazul să le reluăm. Executarea procedurii de analiză determină afişarea în fereastra de rezultate a unui alt tabel:

Univariate Statistics

		Estimate	Standard Error	95% Confidence Interval		Coefficient of Variation	Population Size	Unweighted Count
				Lower	Upper			
Mean	Age of Respondent	46,33	,369	45,60	47,05	,008	1517,000	930

Figura 8.32 – Rezultatele analizei eşantionului

În acest caz nu mai avem de a face cu estimări ale frecvențelor ci cu estimări ale mediilor, variabila analizată fiind o variabilă continuă. Iată că eşantionul nostru estimează o medie a populației de 46,33 de ani cu o eroare standard de 0,36 ani. Media reală a populației se găseşte undeva între 45,60 şi 47,05 ani. Eşantionul este foarte omogen sub aspectul vârstei, după cum rezultă din coeficientul de variație de 0,008, eterogenitatea fiind în jurul valorii de 0,8%. Într-adevăr, eşantionul de 930 de persoane estimează populația țintă de 1517 persoane însă…. cu o eroare standard mare. Însuşi intervalul de încredere de 1,45 ani (47,05 – 45,60) este edificator în acest sens. Este greu să vorbim despre o reprezentativitate a acestui eşantion sub aspectul vârstei subiecților.

Bibliografie

1. **Clocotici, Vasile şi Stan, Aurel. 2000.** *Statistică aplicată în psihologie.* Iaşi : Polirom, 2000.
2. **Curwin, Jon şi Slater, Roger. 1996.** *Quantitative methods for Business Decisions. Fourth Edition.* London : International Thomson Business Press, 1996.
3. **Dancey, Christine şi Reidy, John. 2002.** *Statistics without math for psychology. Second Edition.* Harlow : Prentice Hall, 2002.
4. **Dowdy, Shirley, Weardon, Stanley şi Chilko, Daniel. 2004.** *Statistics for research. Third Edition.* New Jersey : John Wiley and Sons, 2004.
5. **Field, A. 2000.** *Discovering statistics using SPSS for Windows.* London : Sage, 2000.
6. **Field, Andy. 2005.** *Discovering Statistics using SPSS, Second Edition.* London : Sage Publications Ltd., 2005. 100-7619-4451-6.
7. **Jaisingh, Lloyd. 2000.** *Statistics for the Utterly Confused.* New York : McGraw Hill, 2000.
8. **Leech, Nancy L, Barrett, Karen C şi Morgan, George A. 2005.** *SPSS for intermediate statistics. Use and interpretation. Second Edition.* New Jersey : Lawrence Erlbaum Associates, 2005.
9. **Lungu, Ovidiu. 2001.** *Ghid introductiv pentru SPSS 10.0.* Iaşi : Erota, 2001.
10. **Marques de Sa, Joaquim P. 2007.** *Applied Statistics using SPSS, Statistica, Mathlab and R.* New York : Springer, 2007.
11. **Pitariu, Horia Dumitru. 2003.** *Psihodiagnostic - note de curs.* Bucureşti : s.n., 2003.
12. **Popa, Marian. 2003.** *Statistică - note de curs.* Bucureşti : s.n., 2003.
13. **Radu, Ioan, şi alţii. 1993.** *Metodologie psihologică şi analiza datelor.* Cluj-Napoca : Sincron, 1993.
14. **Rotariu, T, şi alţii. 2006.** *Metode statistice aplicate în ştiinţele sociale.* Iaşi : Polirom, 2006.
15. **Sava, Florin Alin. 2004.** *Analiza datelor în cercetarea psihologică. Metode statistice complementare.* Cluj-Napoca : ASCR, 2004.
16. **—. 2000.** Pagina de statistică socială. *Pagina de statistică socială.* [Interactiv] 2000. http://statisticasociala.tripod.com.

17. **SPSS, Inc. 2001.** *SPSS Base 11.0 User's Guide.* Chicago : SPSS Inc, 2001.
18. **—. 2006.** *SPSS Complex Samples 15.0.* Chicago : SPSS Inc, 2006.
19. **System, The Survey.** The Survey System: Sample Size Calculator. *The Survey System.* [Interactiv] http://www.surveysystem.com/sscalc.htm.
20. **Vasilescu, Ilie Puiu. 1992.** *Statistică informatizată pentru ştiinţele despre om.* Bucureşti : Militară, 1992.

În loc de încheiere

Sper că v-am interesat în acest domeniu. După parcurgerea cărții, ați văzut că statistica poate fi chiar plăcută iar concluziile pe care le putem desprinde din analiza datelor, ne conduc către idei pe care, în absența acestei științe, nu le-am putea avea, și, mai ales, nu le-am putea verifica. Bineînțeles, statistica este, întâi de toate, o disciplină matematică. Nimeni însă nu vă cere să fiți un matematician pentru a o putea folosi. Așa cum îmi spunea odată colegul și prietenul meu, Ticu Constantin, „nu e nevoie să știu să construiesc o mașină pentru a o putea conduce" și cu care, la momentul respectiv, nu am fost de acord; acum trebuie să recunosc că a avut dreptate. Rolul matematicienilor este acela de a construi „mașina"; rolul nostru este de a o conduce. Acum, dacă unii șoferi sunt mai buni decât alții… asta e. Nu la fel se întâlnim și în viața reală, pe șosele? Nu este întâmplător faptul că, pionierii psihologiei științifice care au folosit metode statistice (Cattell, Spearman, Thorndike s.a.) au fost, inițial, formați în științe reale precum matematica, fizica, chimia.

În acest volum, m-am oprit la studiul eșantioanelor. Din rațiuni concrete, nu am dorit să merg mai departe, deoarece aș fi fost nevoit să scot un volum de 600-700 de pagini, care, cu siguranță v-ar fi speriat. Nu intenționez însă să mă opresc aici. În volumul al doilea, care va apărea după ce voi obține impresiile dumneavoastră, aș vrea să abordez problema predicției în științele socio-umane, prin metode statistice și nu clarvizioniste, adică metode de regresie liniară și logistică, corelațiile și analiza factorială sau modalități de a vedea dacă datele noastre au ceva în comun între ele, analiza de fidelitate sau cât de stabile pot fi scorurile în timp ori cum putem să facem să nu mai așteptăm un timp pentru a vedea acest lucru, metode prin care putem sesiza dacă există diferențe între grupuri, cu alte cuvinte statistici inferențiale. De asemenea, mi-aș dori să abordez un capitol extrem de interesant și anume acela al analizei de varianță sub diferitele sale aspecte, în care să putem vedea cum una sau mai multe dimensiuni exercită efecte asupra unor alte dimensiuni.

În al treilea volum, aş vrea să abordez problematica scalării multidimensionale, problematica analizei de cluster, analiza Q Sort, metaanaliza, analizele de itemi, analize calitative şi analizele folosind reţelele neuronale.

Demersul este, după cum aţi observat, de la simplu la complex. Aş fi foarte bucuros dacă aş vedea că ridicaţi probleme, propuneţi soluţii, oferiţi sugestii sau, de ce nu, criticaţi anumite aspecte pe care le prezint. Asta ar fi dovada că munca mea nu a fost în zadar. Nu am pretenţia unui expert absolut. Ceea ce doresc, este să partajez experienţa mea cu experienţa altora, din care să avem cu toţii de câştigat. Pentru a realiza acest obiectiv, va trebuie să ne detaşăm de orgolii şi de impresii personale, să conştientizăm că toţi avem încă de învăţat, să avem o minte deschisă şi să dăm ceea ce este mai bun în noi. Este foarte posibil ca, uneori, în ceea ce expun, să nu am dreptate. Este posibil să mă înşel în anumite afirmaţii pe care le fac pe parcursul acestor pagini. Singura modalitate pe care o avem de a ne da seama de acest lucru este să comunicăm. Aştept, prin urmare, reacţia dumneavoastră, ca unici judecători ai cuvintelor mele.

Vă mulţumesc pentru răbdarea pe care aţi avut-o studiind acest volum. Informaţiile prezentate aici nu sunt originale şi nici nu ar avea cum să fie. Ele se bazează pe afirmaţiile unor genii în domeniu, care au avut inspiraţia să le descopere iar eu nu am făcut decât să le prezint. Originală este însă, cred, modalitatea de prezentare. Am încercat să îmbin umorul cu informaţia ştiinţifică, astfel încât cele prezentate aici să fie uşor „digerabile". De asemenea, am urmărit să folosesc cât mai multe exemple practice pentru a ieşi puţin din lumea abstractă a formulelor matematice. Dacă am reuşit sau nu acest lucru, dumneavoastră să-mi spuneţi…